11-30-2017 2:30 PM

Chromatin accessibility dynamics in the Arabidopsis root epidermis and endodermis during cold acclimation

Shawn Hoogstra
The University of Western Ontario

Supervisor
Austin, Ryan S.
The University of Western Ontario Co-Supervisor
Hill, Kathleen A.
The University of Western Ontario

Graduate Program in Biology
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

Abstract

Understanding cell-type specific transcriptional responses to environmental conditions is limited by a lack of knowledge of transcriptional control due to epigenetic dynamics. Additionally, cell-type analyses are limited by difficulties in applying current technologies to single cell-types. A novel DNase-seq protocol and analysis procedure, deemed DNase-DTS, was developed to identify DHSs in the Arabidopsis epidermis and endodermis under control and cold acclimation conditions. Results identified thousands of DHSs within each cell-type and experimental condition. DHSs showed strong association to gene expression, DNA methylation, and histone modifications. *A priori* mapping of existing DNA binding motifs within accessible genes and the cold C-repeat/dehydration responsive element-binding factor pathway resulted in unique motif mapping patterns. In summary, a collection of endodermal and epidermal cold acclimation induced chromatin accessibility sites may be used to understand mechanisms of gene expression and to best design synthetic promoters.

Keywords: chromatin accessibility, epigenetics, Arabidopsis root, DNase hypersensitive sites, epidermis, endodermis, histone modifications, DNA methylation, transcription, cis-regulatory motifs, gene expression, transcription, next generation sequencing

Contents

List of Figures

List of Tables

List of Algorithms

List of Appendices

List of Abbreviations, Symbols, and Nomenclature

ABA	Abscisic acid
*AP*2	APETALA2
bHLH	Basic helix-loop-helix
bZIP	Basic leucine zipper
BLRP	Biotin ligase recognition peptide
CBF	C-repeat binding factor
CRT	C-repeat
COR	Cold-regulated
DDTS	DNase-DTS
DE/*DA*	Differentially expressed differentially accessible
DHS	DNase I hypersensitive site
Dof	DNA-binding-with-one-finger
DRE	Dehydration-responsive element
DTS	Direct to sequencing
FACS	Fluorescence-activated cell sorting
FDR	False discovery rate
FLC	FLOWERING LOCUS C
FPKM	Fragments per kilobase of transcript per million mapped reads
GFP	Green fluorescent protein
GO	Gene ontology

HMG	High mobility group
*ICE*1	Interactor of little elongation complex ELL Subunit 1
INTACT	Isolation of nuclei tagged in specific cell-types
NFR	Nucleosome free regions
NGS	Next generation sequencing
NTF	Nuclei targeting fusion protein
PIC	Preinitiation complex
SBP	SQUAMOSA binding protein
TBP	Transcription binding protein
TF	Transcription factor
TPR	True positive rate
TSS	Transcriptional start site
UTR	Untranslated region

Chapter 1

Introduction

1.1 Transcriptional regulation in Arabidopsis

A key question at the core of genetics is how, where, and when a gene is transcribed? How does the information encoded in a simple DNA sequence of 4 base pairs result in the complex expression of RNA and proteins forming diverse cellular functions? While the study of gene expression has been at the forefront of biological research, a complete answer to these questions still remains. Despite the basics of transcription and the *central dogma* of biology being understood, an understanding into transcription's complex regulation has been difficult. Indeed, while countless transcriptomic studies have been performed in the last 20 years, researchers only partly understand the intricacies of transcriptional regulation.

To begin with, transcription has been difficult to understand due to transcription being a highly complex system involving the interaction of many components temporally and spatially. Notably, transcription factor (TF) binding, DNA methylation, chromatin accessibility, and histone modifications all interact and influence transcription (Arnone and Davidson (1997); Barlow (1993); Zhang et al. (2012a); Karlić et al. (2010); Siegfried et al. (1999)). In order for an organism to efficiently develop and respond to environmental conditions, it requires all those components to interact in very tightly controlled ways. For instance, while TF binding mainly involves the interaction between specific DNA sequences and TFs, other processes can inhibit or enhance this interaction. Specifically, high amounts of DNA methylation can reduce tran-

scriptional output by interfering with TF binding (Siegfried et al. (1999); Vining et al. (2012); Thurman et al. (2012); Jones et al. (1998); Klose and Bird (2006)). Likewise, histone modifications are linked to transcriptional output. However, histone modification's link to transcription may be indirect due its influence on the overall chromatin structure rather than interfering with TF binding directly (Kouzarides (2007); Zhang et al. (2007); Ernst et al. (2011); Thurman et al. (2012); Bernstein et al. (2002)). This being said, transcriptional output is certainly dependent on chromatin structure. Chromatin structure is precisely defined by overall accessibility to TFs which will directly influence gene expression. A full description of these processes and how they influence transcription is discussed in the following subsections.

1.1.1 Transcriptional activation

In order for a gene to be expressed it requires, first and foremost, the binding of the main transcriptional preinitiation complex. The preinitiation complex (PIC) is a large assembly of proteins involving the interaction of an RNA polymerase and several general TFs. Once this interaction is formed near the transcriptional start site (TSS), the complex transcribes DNA into messenger RNA (Allen and Taatjes (2015); Darnell (2013)). Furthermore, additional TFs will bind to nearby DNA regulatory elements and alter, either as an activator or as a repressor, transcription through changing how much, where, and when a gene should be expressed.

Regulatory elements are DNA sequences which TFs recognize in order to bind to DNA and control expression. These may be enhancers, enhancing the expression of the associated gene or repressors, repressing the expression of the associated gene. These elements may contain known binding sites, called motifs, with which the TFs specifically bind. Over 1,500 characterized TFs are known in the Arabidopsis genome which control gene expression in a cell-type and condition dependent manner (Riechmann (2002); Kaul et al. (2000)). In order to bind to their DNA sequences, many TFs require activation through ligand binding, signal cascades, or interaction with other proteins. TFs have been found to recruit chromatin remodelling complexes which alter the surrounding chromatin structure. By recruiting these complexes and

altering the surrounding chromatin structure, TFs modify gene expression by changing how accessible chromatin is to other TFs and the basal transcriptional machinery (Li et al. (2001); Yudkovsky et al. (1999)).

An additional layer of control, the epigenetic layer, highly influences and controls transcription spatially and temporally through enhancing or inhibiting TF binding (Kouzarides (2007)). Epigenetics is defined as any heritable change in transcription not explained through changes in the DNA sequence. The epigenetic layer includes histone modifications, DNA methylation, and chromatin accessibility, all separate from the four base pairs (Siegfried et al. (1999); Kouzarides (2007)). In summary, transcriptional activation is a complex process involving the interaction of the DNA sequence, TF binding, and the epigenetic layer.

1.1.2 Transcription factor families

The *Arabidopsis thaliana* genome contains roughly 1,572 TFs that belong to 45 different TF families based on their sequence and functional similarities (Riechmann et al. (2000); Riechmann (2002); Kaul et al. (2000)). TF families are not only structurally similar but recognize and bind related motifs (Jakoby et al. (2002); Pabo and Sauer (1992); Weirauch and Hughes (2011); Weirauch et al. (2014)). There are many TF families for example, the basic helix-loop-helix (bHLH), Myb/SANT, C_2H_2 zinc fingers, homeodomain, and basic leucine zipper (bZIP). In addition, plants contain several plant specific TF families including the MADS box, APETELA2 (AP2), DNA-binding-with-one-finger (Dof), Whirly, B3, WRKY, NAC, and SQUAMOSA binding proteins (SBP) (Weirauch and Hughes (2011)). Notably, the largest family of TFs, the AP2 family, is critical in abiotic stress response including cold and drought (Dietz et al. (2010); Stockinger et al. (1997); Fowler and Thomashow (2002)).

In Arabidopsis, 414 motifs representing 666 TF genes have been identified and characterized (Weirauch et al. (2014)). Through identifying these motifs and mapping their locations across the genome, one may begin to understand where and when these TFs are binding. Thus, identifying and mapping these motifs, using specialized software (e.g. Austin et al. (2016)), is

critical for understanding how genes are expressed temporally and spatially.

1.1.3 Methods of transcriptional quantification

In order to understand why and how transcription occurs, one must first have a complete understanding of an organism's transcriptome. The transcriptome is the identity but also the quantity of all the transcripts in any tissue, cell, or organism. The main goal of understanding the transcriptome is to identify and quantify the expression of each gene throughout development, under certain conditions, and in certain tissues or cell-types. Through understanding what is being transcribed and where, researchers gain a greater understanding of how an organism develops or responds to internal and external conditions.

Currently, RNA-seq is the most common method for quantifying transcription levels (Weber et al. (2007); Marioni et al. (2008); Wang et al. (2009)). RNA-seq is a high-throughput sequencing method to accurately identify and quantify gene expression across many organisms and tissue types. RNA-seq takes cDNA fragments, converted from RNA, and sequences the cDNA fragments in a high-throughput method (Weber et al. (2007); Marioni et al. (2008); Wang et al. (2009)). Afterwards, sequenced reads are mapped to the reference genome and information containing the transcript structure and expression level are analyzed. Transcript abundance is subsequently used to identify differentially-expressed genes between samples or to integrate transcript abundance with other data sources. To conclude, RNA-seq has become a very popular tool for identifying and characterizing the transcriptome within various species, tissues, and cell-types (Weber et al. (2007); Marioni et al. (2008); Wang et al. (2009); Zeisel et al. (2015); Li et al. (2016)). As a result, a huge amount of sequencing data have been generated and are publicly available for researchers to use and integrate in their various experiments (e.g. Kanz et al. (2005); Benson et al. (2012); Barrett et al. (2012)).

1.2 Epigenetics

Transcription factor binding to the DNA sequence will only explain part of how, why, and where transcription takes place. Each cell in an organism contains the exact same genetic code in which only a limited amount of information can be transferred. Thus, to fully understand transcription a complete understanding of influences above the DNA sequence is required. Specifically, to create a complete picture of transcription, researchers require an understanding of epigenetic's control on gene expression.

Epigenetics includes several layers including DNA methylation, histone modifications, and chromatin remodelling or packaging. Together these layers, in addition to the DNA sequence and TFs, interact and modify each other resulting in specific control of gene expression (Siegfried et al. (1999); Kouzarides (2007); Felsenfeld (1992); Yuan et al. (2005)). Indeed, TF binding in combination with epigenetics is a tightly controlled and interconnected process that leads to specific control over gene expression developmentally and in response to external conditions (Feil and Fraga (2012); Kiefer (2007)). The subsequent subsections detail the relevant background and current understanding of epigenetics and its effect on transcription and TF binding.

1.2.1 Chromatin accessibility

DNA is not simply a linear strand with no structure, packaging, or organization. Instead, DNA forms a tightly regulated complex with proteins packaging the DNA into a highly compact and dense structure called chromatin. By forming this complex, DNA is protected from external damage while enabling a greater control over gene expression. (Felsenfeld (1992); Workman and Kingston (1998); Yuan et al. (2005)). To form the DNA-protein complex, a core of proteins called histones tightly bind and wrap DNA to form a nucleosome. A nucleosome is composed of two copies of histones H2A, H2B, H3, and H4, making up a histone octamer (Van Holde (2012)). The nucleosome is the basic unit of chromatin that packages DNA with further levels

of compaction leading to highly dense chromatin.

The chromatin structure is highly flexible and dynamic constantly undergoing alterations. DNA may change from a highly compacted state wrapped around nucleosomes to a state not bound to any nucleosomes. In a nucleosome depleted state, DNA is highly accessible for TF binding and interactions with other regulatory proteins. However, in a nucleosome bound state, TFs are inhibited from binding to their respective cis-regulatory elements (Felsenfeld (1992); Workman and Kingston (1998); Yuan et al. (2005)). As a result, nucleosome-depleted regions are deemed accessible or open, while nucleosome bound DNA regions are deemed closed or inaccessible.

However, rather than being a strict case of open or closed, chromatin is instead a continuous spectrum from closed to open (Zhang et al. (2012a); Liu et al. (2017); He et al. (2012)). This spectrum or degree of accessibility is defined as DNA accessibility. Chromatin may be in any state from DNA tightly wrapped around nucleosomes to a state where no nucleosomes are present. One may identify these open sites and define their accessibility using an enzyme known as DNase I, an endonuclease (Wu et al. (1979)). Hence, these nucleosome-depleted regions are known as DNase I hypersensitive sites (DHSs) due to their high sensitivity to DNase I digestion. Hence, using this enzyme, detailed information into the locations and accessibility of DHSs may be obtained.

In order to obtain a complete picture of transcription, an understanding of DNA accessibility is required. Specifically, DNA accessibility has been linked to transcription through TF binding interference (Guertin and Lis (2013); John et al. (2011); Bell et al. (2011)). In a condensed or closed chromatin state, the ability of transcriptional components to bind to their DNA sequences is interfered with. Furthermore, in an open and accessible state, transcriptional components may bind to their respective DNA sequences (Figure 1.1; Felsenfeld (1992); Workman and Kingston (1998); Yuan et al. (2005)). As a result, open chromatin in gene promoters is associated with active transcription while closed chromatin in gene promoters is associated with inactive transcription (Zhang et al. (2012a); Boyle et al. (2008a); Song et al. (2011)).

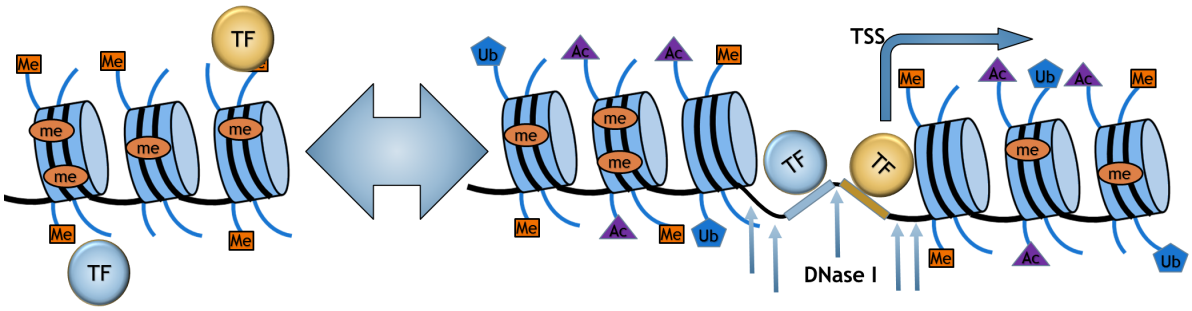


Figure 1.1: **Gene expression requires the coordination of histone modifications, DNA methylation, chromatin accessibility, and TF binding.** Depicted is a cartoon image of all these factors interacting to drive expression of a downstream gene. Closed chromatin with distinct modifications are shown on the left. Accessible DNA is shown on the right with distinct chromatin structure and modifications surrounding. TF binding is inhibited by closed chromatin while enabled in accessible DNA. Orange squares represent histone methylation, orange ovals represent DNA methylation, blue pentagons represent histone ubiquitination, and purple triangles represent histone acetylation. DNase I preferentially digests within accessible chromatin.

In addition to the presence of accessible chromatin affecting transcription, the degree of DNA accessibility significantly impacts transcriptional output. Previous research identified that genes with highly sensitive promoter DHSs have higher gene expression than genes with less sensitive DHSs (Zhang et al. (2012a); Natarajan et al. (2012)). In other words, genes more accessible to TF binding will be on average more highly expressed.

However, even though genes may have highly accessible promoters, they may not be highly expressed or expressed at all. As mentioned earlier, for a gene to be expressed it requires TF binding. A gene may be accessible to TF binding but lack TFs bound to its regulatory elements leading to no transcription. These genes are in what is called a transcriptionally poised state. They are accessible and poised for TF binding and therefore expression, but are not necessarily actively transcribed and bound to TFs (Gross and Garrard (1988); Sullivan et al. (2014, 2015)). Such genes are often controlled throughout development and under specific environmental conditions (Sullivan et al. (2014)). These accessible sites require their binding TFs to be expressed or activated prior to binding. Sullivan et al. (2014, 2015) suggested this may be due to the high amount of energy required for changing chromatin states. Therefore, to reduce energy usage, genes may remain highly accessible for TF binding. Furthermore, organisms like plants are required to respond to conditions quickly and altering chromatin states is quite time intensive. As a result, maintaining an active open chromatin state may significantly reduce response times in response to rapidly changing environmental conditions (Raser and O'Shea (2004)).

Due to chromatin's association with TF binding and gene expression, DHSs are linked with genomic features like transcriptional start sites (TSSs), enhancers, suppressors, and transcription factor binding sites. In general, cis-regulatory regions, active regulatory regions, or any DNA regions requiring protein binding or interaction are associated with accessible chromatin (Gross and Garrard (1988); Natarajan et al. (2012); Boyle et al. (2008a); Heintzman et al. (2007)). Therefore, identifying DHSs will aid in predicting new gene regulatory regions or "functional" regions throughout genomes. Predicting new regulatory regions, like enhancers,

is of particular importance in Arabidopsis as they have been very difficult to identify using existing techniques and there are many to identify (Zhu et al. (2015)). Lastly, in addition to predicting new regulatory regions, identifying DHSs also aids in identifying the specific TF binding sequences or motifs of those regions.

DNA accessibility is important for understanding transcription throughout an organism, in specific cell-types, throughout plant development, and in response to environmental stimuli (John et al. (2011); Sullivan et al. (2014); Pajoro et al. (2014); Song et al. (2011)). In order to understand how organisms develop, how cell-types form, and how organisms respond to stressful conditions, it is important to understand how those processes are affected through chromatin dynamics. Particularly, in Arabidopsis, DNA accessibility has been mapped and shown to be important in leaf and flower development, in seedling development, in root development, and in response to heat and light (Pajoro et al. (2014); Zhang et al. (2012b); Sullivan et al. (2014); Cumbie et al. (2015); Liu et al. (2017)). In general, DNA accessibility was found to be involved in the regulation of transcription for Arabidopsis developmental and stress responsive genes. Developmental and environmental response pathways induce changes in chromatin structure leading to alterations in transcriptional output (Jiang (2015); Chinnusamy and Zhu (2009); Luo et al. (2012)).

However, while whole tissue and organism samples have been well studied, cell-type specific studies in Arabidopsis are lacking. This may be due to the difficulty in isolating and accumulating enough nuclei using current protocols, especially in very difficult tissues like the root (Cumbie et al. (2015)). Despite this, several studies attempted and have shown the importance of chromatin dynamics in cell-type development (Stergachis et al. (2013); Song et al. (2011); Costa and Shaw (2006)). One study focused on identifying how the root epidermis develops in alternating patterns of epidermal cells and hairs cells. Their results found the epidermal alternating pattern was due to gene expression changes as a result of chromatin state alterations around one gene locus, the *GL2* locus (Costa and Shaw (2006)). As a result, in order to understand cell-type development and formation, it is important to understand tran-

scriptional alterations due to epigenetic changes. It is important to not only understand the epigenetic differences between cell-types, but also how chromatin alterations arise and affect transcription.

1.2.2 Chromatin remodelling, histone modifications, and DNA methylation

DNA accessibility and chromatin remodelling, or modifying the chromatin structure, is controlled through many processes affecting the DNA-protein interaction of nucleosomes. One process that remodels chromatin involves various nucleosome remodelling complexes using ATP-hydrolysis to establish accessible chromatin through two distinct processes. (Clapier and Cairns (2009); Becker and Hörz (2002); Bell et al. (2011)). The first process involves the the SWI/SNF complexes which slide nucleosomes along DNA into adjacent DNA regions allowing them to become accessible or closed (Becker and Hörz (2002); Brehm et al. (2000); Bell et al. (2011)). This can make it difficult to identify DHSs as nucleosome sliding is considerably dynamic and fluid. The second process involves remodelling complexes evicting or removing nucleosomes completely to create accessible DNA (Becker and Hörz (2002); Phelan et al. (2000); Bell et al. (2011)). Through these two processes, nucleosome sliding or eviction, chromatin remodelers allow DNA to become accessible to TFs.

In addition to chromatin remodelling, chemical modifications to histone components will aid in altering DNA accessibility by either recruiting other remodelling factors or affecting nucleosome stability (Bell et al. (2011)). First, many documented proteins bind to chemical modifications on histone tails that proceed to alter surrounding chromatin (Taverna et al. (2007)). As a result, histone modifications often precede changes to the overall chromatin structure. For example, H3K27me3 was found to recruit Polycomb proteins that proceed to close and compact the surrounding chromatin (Bell et al. (2011); Francis et al. (2004)). Second, histone modifications are thought to destabilize histone/DNA interactions in nucleosomes or between adjacent nucleosomes. Due to this destabilization, DNA then becomes more accessible. For

instance, H4K16ac was identified directly affecting the interactions between adjacent nucleosomes allowing DNA to become accessible (Shogren-Knaak et al. (2006); Bell et al. (2011)).

There are too many histone modifications to be discussed within the framework of this thesis, hence, only those used in the current work are discussed. Two modifications, H3K4me3 and H3K27me3, were selected as they are well characterized in the context of chromatin accessibility. H3K4me3 is found linked with open chromatin and active transcription while H3K27me3 has been linked to closed chromatin and inactive transcription (Kouzarides (2007); Zhang et al. (2007); Ernst et al. (2011); Thurman et al. (2012); Bernstein et al. (2002)). However, while histone modifications are associated with regions of open and closed chromatin, accessible DNA regions are depleted of nucleosomes and therefore depleted of the majority of histone modifications (Figure 1.1; Zhang et al. (2012a)). Despite this, regions surrounding DHSs are associated with distinct epigenetic markers due to DHSs being flanked by highly positioned nucleosomes or strongly phased nucleosome arrays (Wu et al. (2014); Radman-Livaja and Rando (2010)). These highly positioned nucleosomes can lead to very distinct epigenetic spikes in histone modifications surrounding DHSs (Zhang et al. (2012a)).

While DNA methylation may not be directly involved with chromatin modifications or alterations, DNA methylation is associated and linked with chromatin accessibility patterns (Figure 1.1). For instance, DHSs are often hypomethylated or associated with a lack of DNA methylation. In the context of transcription, hypomethylation occurring alongside DHSs makes biological sense as both are associated with active transcription (Zhang et al. (2012a); Sullivan et al. (2014); Lister et al. (2009)). However, while distinct methylation patterns are associated with DHSs, it is unknown whether DNA methylation precedes or follows DNA accessibility changes or how they influence one another (Jones et al. (1998); Wade et al. (1999); Chodavarapu et al. (2010)). CpG, CHG, and CHH are the three types of DNA methylation within plants where H is an adenine, cytosine, or thymine. While CpG methylation is biologically universal, CHG and CHH methylation are mostly unique to plants and fungi (Suzuki and Bird (2008); Lister et al. (2009))

1.3 DNase I hypersensitive site identification

A challenge within genomics is to accurately and efficiently identify DHSs on a large scale throughout entire genomes in a high-throughput manner. To date, several protocols and strategies have been developed to accomplish efficient identification of DHSs through similar but differing approaches. The majority of DHS identification protocols rely on the basic property that DHSs are sensitive to digestion by an enzyme known as DNase I (Wu et al. (1979)). After DNase I digestion, each protocol identifies DHSs through their own various computational analysis strategies and techniques. The three main strategies deployed in DHS identification are Southern blotting, DNase-chip,and DNase-seq (Wu et al. (1979); Crawford et al. (2006); Boyle et al. (2008b); Hesselberth et al. (2009))

1.3.1 Current DHS identification strategies

Previous DHS identification protocols involved DNase I digestion followed by Southern blotting (Wu (1980)). Any DHS digested would appear as distinct smearing patterns and bands on agarose gels. Despite this protocol leading to the identification and characterization of many DHSs, it was time consuming, limited in scale and output, and limited in detailed results (Wu (1980); Keene et al. (1981)). Therefore, to deploy DHS identification on a large scale, new techniques using high-throughput methods were needed.

As a result, DNase-chip was developed to replace Southern blotting as a high-throughput DHS identification method (Crawford et al. (2006)). DNase-chip involves the same basic process of digesting nuclei with DNase I and isolating DNA for analysis. In this method, DNA is slightly digested with DNase I so DHSs are cut only once or twice per fragment. This results in DNA fragments whose ends are the sites of DNase I digestion. The ends of these fragments are isolated, run on microarrays, and analyzed through existing microarray analysis tools. Genomic regions containing a high enrichment of these fragment ends are identified as DHSs. Thus, DNase-chip allowed for the quick identification of DHSs on a large scale in

a high-throughput manner. However, DNase-chip is still low resolution, lacks genome wide coverage, and is highly dependent on the microarray used. Despite this, microarray analysis of DHSs is still widely used as its main advantage is ease of use and quite sophisticated analysis tools.

A recent improvement in DHS identification methods was the development of DNase-seq (Boyle et al. (2008a); Hesselberth et al. (2009)). DNase-seq overcomes and significantly improves upon the issues present in previous DHS technologies allowing for single base pair resolution of DHSs genome wide. As with previous strategies, DNase-seq involves slightly digesting DNA with DNase I so the ends of each DNA fragment represent one DNase I cut site. However, rather than isolating the ends of DNase I digested fragments and hybridizing them to an array, DNase-seq sequences the ends of DNase I digested fragments. DNase-seq, therefore, produces a genome-wide picture of chromatin not limited to the information on a microarray. Similar to DNase-chip, DNase-seq analysis involves identifying regions of high fragment enrichment. Despite improving upon microarray technologies considerably, DNase-seq has somewhat immature and difficult to use analysis tools and so many studies prefer to use DNase-chip. Finally, DNase-seq is significantly more expensive than microarray technologies but with lowering sequencing costs, DNase-seq is slowly becoming the preferred option.

A limiting issue with DNase-seq or any current molecular protocol is the difficulty in working with single cell-types. Current protocols involve the requirement of millions of nuclei which are difficult to produce for many organisms and cell-types (Boyle et al. (2008a)). In addition, DNase-seq protocols are time consuming and involve extensive biochemistry making working with resistant tissues difficult (Cumbie et al. (2015)). For example, it is difficult to obtain enough high quality nuclei even in the entire Arabidopsis root using existing DNase-seq protocols, as it is a difficult and uncooperative tissue (Cumbie et al. (2015); Filichkin and Megraw (2017)). Existing nuclei isolation protocols cannot obtain the required abundance or purity of nuclei needed from the Arabidopsis root. Nuclei isolations from this tissue are often plagued with high amounts of cellular debris which purify with the nuclei (Filichkin and

Megraw (2017)). To move DNase-seq to a cell-type resolution, steps toward more efficient nuclei isolation techniques and DNase-seq protocols are required.

A second issue holding DNase-seq back is the use of agarose gels and agarose gel plugs. To prevent mechanical shearing of DNA, agarose gel plugs are required in DNase-seq protocols (Boyle et al. (2008a); Crawford et al. (2006)). In order to obtain enough DNA in difficult tissues like the Arabidopsis root, a large amount of agarose is required for embedding digested nuclei (Filichkin and Megraw (2017)). Furthermore, multiple agarose plugs are required, making the protocols time consuming and inefficient. In addition, the agarose analysis step introduces considerable human error as technicians need to run and interpret gel smears to assess if samples are optimally digested. The agarose analysis is done over a DNase I dilution series in order to select the optimally digested sample for sequencing. Two recent papers by, Filichkin and Megraw (2017) and Cumbie et al. (2015), overcame this issue by removing agarose gel plugs completely. Indeed, Cumbie et al. (2015) produced enough DNA to perform DNase-seq on the Arabidopsis root through their method. However, agarose analysis and large amounts of nuclei are still required which cannot be obtained from Arabidopsis root cell-types. Lastly, despite existing protocols overcoming many issues, current protocols are still time consuming, involve inefficient fragment enrichment steps, extensive library preparation protocols, and blunt end polishing. Hence, a new protocol and analysis strategy requires development so DNase-seq may be performed on small samples at a single cell-type level in the Arabidopsis root.

1.3.2 DNase direct to sequencing

DNase-DTS (direct to sequencing) is a protocol developed in the Austin lab to overcome several issues surrounding existing DNase-seq protocols while improving upon and introducing new analysis techniques and quality control mechanisms (Unpublished, Figure 1.2). Firstly, DNase-DTS removes agarose gel plugs and improves upon existing DNase-seq nuclei isolation protocols by incorporating the INTACT protocol (Deal and Henikoff (2011)). INTACT allows for high yield and pure nuclei extractions in a relatively easy and time efficient proto-

col. Secondly, DNase-DTS improves upon time consuming steps of isolating and enriching DNase I cut ends by avoiding those steps. Instead DNase-DTS digests nuclei with a large amount of DNase I in order that accessible regions are heavily digested. Afterwards, DNA is isolated from nuclei and sent directly for sequencing. The agarose analysis step is skipped entirely and assessment of DNase I digestion and selection of the optimally digested sample is performed post-sequencing. Instead of several days of work, as is done with previous DNase-seq protocols, DNase-DTS is accomplished in the span of a day. However, this changes the basic fundamentals of DNase-seq. Rather than sequencing DHSs and identifying regions of sequence enrichment, DNase-DTS sequences closed chromatin as the DHSs are heavily digested. In addition, the analysis of DNase-DTS involves the identification of regions lacking sequencing as opposed to enriched regions. Due to the novel data generated, new analysis tools and pipelines needed to be developed in order to identify DHSs. One of the goals of this work was to develop new bioinformatic tools to analyze this new form of data.

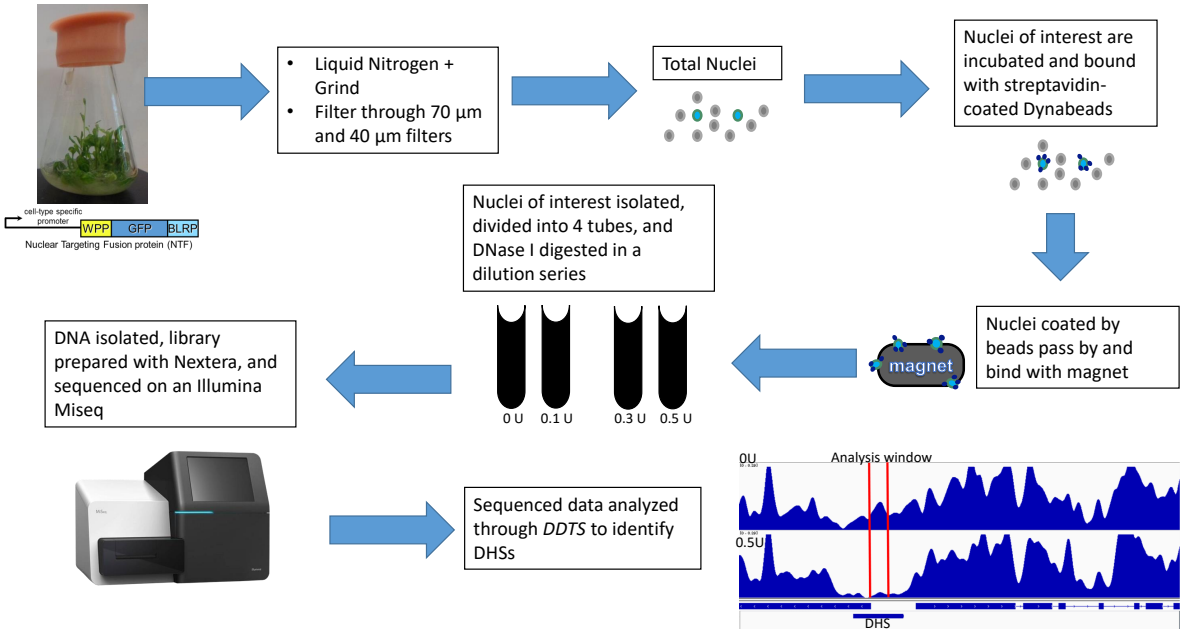


Figure 1.2: **Overview of DNase-DTS protocol and analysis procedure.** Transgenic Arabidopsis containing a nuclei targeted fusion protein with a nuclear envelope targeting signal called WPP, a green fluorescent protein (GFP), and a biotin ligase recognition peptide (BLRP), are ground in liquid nitrogen and filtered through 70 µm and 40 µm filters. Total nuclei are isolated and incubated with streptavidin-coated Dynabeads. Nuclei with a biotin signal bind to the Dynabeads and are isolated through the interaction of a magnet pulling out the magnetic Dynabeads. Nuclei are divided into four separate tubes and digested over a DNase I dilution series. Isolated DNA is sequenced using an Illumina MiSeq. Generated data are then analyzed through custom software called *DNase-DTS (DDTS)* and DHSs are identified.

1.4 Advantages of identifying accessible chromatin

The identification of cell-type specific DHSs considerably increases the knowledge of basic biological processes. For instance, chromatin dynamics may be integrated with transcription to further understand chromatin's role in influencing gene expression and TF binding. In addition, research into cell-type DHSs increases the understanding of how chromatin dynamics influence cell-type formation and development. Through extension of this research into stress and acclimation response, a model of how chromatin dynamics initiate cellular responses to environmental conditions may be made. Lastly, DHSs can be integrated with epigenetic data to identify how they influence one another and interact to achieve proper cell function.

A more practical advantage of identifying accessible chromatin regions, particularly in Arabidopsis, is for the creation of new biotechnological applications and transgenic plants. Through an understanding of chromatin dynamics more intelligently designed synthetic promoters may be designed to take advantage of chromatin dynamics influence over gene expression. Researchers would gain a greater control over where, when, and how genes are expressed. Extension of chromatin research to cell-type development and acclimation responses would allow researchers to develop synthetic promoters expressed in certain cell-types and under certain conditions. Synthetic genes may be designed to be accessible under certain conditions enabling expression of that gene under those conditions. Currently, synthetic genes are typically designed with constitutive promoters resulting in gene expression across all tissues constitutively. However, many transgenes require expression only in certain cell-types under certain conditions for them to be effective. Additionally, in markets within Europe where GMOs are unwanted, not expressing transgenes in the fruit is advantageous. For example, it is attractive for those markets to express an insect resistance gene in leaf tissues but not in the fruit or vegetable that the consumer would consume. Lastly, while transgenic expression of genes may give several benefits, there are often disadvantages. Expression of cold-regulated genes (*COR*), while increasing cold resistance, often reduces the growth rate of plants (Jaglo-Ottosen et al. (1998); Gilmour et al. (2004)). Expression of the transgene specifically under cold ac-

climation or other stressful conditions would be significantly advantageous in such cases. In summary, utilization of chromatin dynamics in biotechnological applications hopes to increase our understanding and control over transgene expression. Combining chromatin dynamics with TF binding, motif sites, histone modifications, and DNA methylation may allow genes to be controlled in a very specific fashion temporally and spatially.

1.5 Cell-type specific root genomics

A large proportion of genomic studies involve the collection of entire tissues from organisms on which the experiment is performed. While these studies have led to many discoveries, they miss important data within separate cell-types. Single cell-type analyses improve not only an understanding into general biological problems like transcriptional control, but also aid in understanding cell-type development and formation (Trapnell (2015)). For instance, cell-type specific studies enable researchers to identify how heterogeneous cell-types form and develop from identical DNA sequences. Likewise, integration with transcriptional data enable a further understanding into how cell-types form while improving our understanding into the basic model of transcription. In fact, understanding cell-type chromatin alterations will lead to the same information on the processes and proteins that lead to chromatin alterations obtained from an entire tissue study. By performing these studies on individual cell-types we gain additional information on cell-type development and transcriptional control which could not be obtained through studies on entire tissues.

Analyses involving whole tissues suffer from issues arising from using heterogeneous samples (Trapnell (2015)). Namely, the output of any experiment involving whole tissues will be the average of the population of cells. Data specific to single cell-types may not be detected as it would be masked by the combined results of other cell-types. Observations from distinct cell-types are consumed by the heterogeneity of the population. A single cell-type analysis reduces these cumulative effects and allows observations of higher resolution. For instance,

results identified from entire root samples would be attributed to the entire tissue. However, those observations may result from a strong signal coming from the epidermal layer masking results from other cell-types (Trapnell (2015)). Therefore, in order to understand basics of biology, like transcription and translation, a greater understanding of how such processes work at the cell-type level is necessary.

1.5.1 The Arabidopsis root

Arabidopsis is a great model organism for use in cell-type specific analyses. Specifically, the Arabidopsis root is of particular use in genomics due to its radial design and highly specialized or distinct cell layers (Birnbaum et al. (2003); Benfey and Schiefelbein (1994)). From the outermost layer inwardly the layers of the root are: the epidermis, cortex, endodermis, and stele (Dolan et al. (1993)). All cell layers, except the stele, are a single cell in depth allowing genomic experiments to be performed in a very targeted manner (Dolan et al. (1993)).

The epidermis is the outermost layer and is the first responder to any external environmental changes. The epidermis protects the root from external threats while acting as the first tissue uptaking chemicals, nutrients, and water. By increasing its surface area, through tubular extensions called root hairs, the epidermis uptakes all necessary nutrients and water for the plant. The cortex is the next layer and accumulates starch as well as aiding in gas exchange and oxygen storage. Continuing further within the root, the endodermis is a critical cell layer serving as the absorbing region of the root while controlling the flow between the outermost layers and the stele. Within the endodermis is a tight barrier called the casparian strip tightly controlling the flow of solutes and water into the vascular stele. The last cell-type in the Arabidopsis root is the stele or vascular tissue which is critical for transporting solutes and water throughout the plant. The stele includes the pericycle, phloem, and xylem. Specifically, the xylem is critical for transporting water from the root to the rest of the plant, while the phloem is important for transporting nutrients and signalling molecules (Esau (1977); Enstone et al. (2002)).

In summary, the Arabidopsis root contains very distinct and highly specialized cell layers

making it a great model organism to understand how cell-types develop and form. Despite cell-type specific differences in the Arabidopsis root being well understood, the specific mechanisms of their development is not. For instance, how do the epidermis and endodermis cell layers of the Arabidopsis root develop to be so highly specialized? One mechanism, chromatin remodelling, is one answer to this question and is the main focus of this work.

1.5.2 Root epidermis and endodermis

This thesis focused on the Arabidopsis root epidermal and endodermal cell layers for several reasons. Transgenic Arabidopsis lines to isolate root epidermis and endodermis nuclei have been developed in the Austin lab. As was mentioned, the epidermal root is the first layer exposed to any form of environmental changes and is the first responder. Particularly, the root is exposed to distinct changes in the temperature of the soil or the liquid media within which they are grown. Similarly, the endodermis layer is expected to be critical for cold stress as it controls the flow of water into the vascular tissue, making it critical for drought response in the root (Esau (1977); Enstone et al. (2002); Kiegle et al. (2000); Henry et al. (2012)). For cold stress and acclimation this is important, as a significant amount of cold damage is from freezing induced dehydration (Steponkus and Webb (1992)). Lastly, the root shows distinct changes in gene expression under cold stress with only 14% of its transcriptome changes being shared with changes in the leaves (Kreps et al. (2002)). Thus, the root appears to drastically respond to cold conditions in a distinct manner and research into root cold response will produce unique results. In summary, researching various biological aspects of root epidermis and endodermis development will increase understanding into chromatin dynamic's influence on cell-type development and cold acclimation.

1.5.3 Cell-type specific isolation methods

A challenge with cell-type analyses is the isolation of pure distinct cell-types. The use of cultured cell lines has simplified this problem for a lot of organismal tissues. However, for many

organisms and tissues, obtaining pure single cell-types can be quite challenging. One process to overcome these challenges is fluorescence-activated cell sorting (FACS) (Bonner et al. (1972)). FACS passes cells, one at a time, through a laser beam to identify and separate those that have been fluorescently labelled with a fluorescent marker protein. However, this process requires expensive equipment, is difficult to operate, uses harsh chemicals, and provides low quality nuclei (Deal and Henikoff (2011)).

A new protocol, isolation of nuclei tagged in specific cell-types (INTACT), was developed to overcome many of the FACS issues (Deal and Henikoff (2010, 2011)). In this method, plants are transformed with a transgene containing a nuclear envelope targeting signal, a green fluorescent protein, and a biotin ligase recognition peptide (Deal and Henikoff (2010)). Expression of the fusion protein using a cell-type specific promoter and a nuclear envelope targeting signal inserts the folded protein containing the biotin ligase recognition peptide into the nuclei of interest. A biotin ligase is co-transformed with the previous transgene in order for a biotin marker to be added to the biotin recognition peptide. As a result, the nuclei become biotinylated as the biotin ligase recognition peptides became tagged with a biotin label. Nuclei of interest are now easily isolated from the organism using the interaction of the biotin marker with streptavidin coated magnetic beads. The INTACT protocol thus allows nuclei to be isolated relatively easily and with high yield and purity (Deal and Henikoff (2010)).

1.6 Acclimation and stress response

Plants have adapted many processes to deal with the fact they are sessile organisms and must cope with environmental conditions. For agriculture purposes, research into plant environmental adaptations is critical as extreme environmental conditions significantly impact plant growth and yield. In addition, due to climate change affecting crop growth and yield, and an ever increasing world population, the agriculture industry is under considerable pressure to produce enough food for the world (Godfray et al. (2010); Singh (2012); Sinha et al. (2015)). Thus,

in order to develop crops with high yield and resistance to extreme environmental conditions, a clear understanding of the mechanisms through which plants respond to adverse conditions will be critical. Specifically, understanding how crops acclimate to stressful conditions will enable future GMOs and agriculture practices that could significantly improve crop survivability and yield while being exposed to stressful conditions.

Within temperate regions, cold is a major factor significantly impacting the survivability and growth of crops (Thakur et al. (2010); Kasuga et al. (1999); Sinha et al. (2015)). Cold stress includes two types: freezing stress, characterized as less than 0°C; and chilling stress, characterized as less than 20°C. Freezing stress and chilling stress significantly affect the yield and survivability of many agriculturally important crops including rice, cotton, maize, and soybean (Board et al. (1980); Thakur et al. (2010); Kargiotidou et al. (2010); Rymen et al. (2007); Sionit et al. (1987)). However, acclimating plants to cold conditions, whereby they become accustomed to cold conditions, significantly improves their ability to survive cold conditions (Wanner and Junttila (1999); Guy et al. (1985)). Many plants do not carry the ability to cold acclimate while others do (Chinnusamy et al. (2007); Wanner and Junttila (1999)). Therefore, in order to enhance cold survivability of crops, a proper understanding of how cold acclimation occurs will be necessary. Understanding cold acclimation will aid in enhancing existing cold acclimation but also in giving crops the ability to cold acclimate. It is interesting to note cold and freezing stress also impart severe dehydration on plants. As a result, understanding the effects of cold on plants will not only enable an understanding of how plants respond to cold, but also reflect insights of how plants respond to dehydration (Steponkus and Webb (1992); Fowler and Thomashow (2002); Steponkus et al. (1998)).

1.6.1 Cold stress and acclimation pathways

Plants respond extensively to cold stress and acclimation through the activation of many genes and pathways. In *Arabidopsis* the exact number of cold regulated genes is unknown but, depending on the study, between 4% and 14% of all genes are cold-regulated genes (Hannah

et al. (2005)). One of the most studied cold stress and acclimation pathways is called the C-repeat binding factor (CBF) pathway. The CBF pathway controls many of the downstream cold-regulated genes upregulated during cold stress and acclimation. The most studied proteins within this pathway are ICE1 (Interactor of little elongation complex ELL Subunit 1) and CBF transcription factors (Chinnusamy et al. (2003); Medina et al. (2011)). The three CBF TFs: CBF1, CBF2, and CBF3, are part of the AP2 transcriptional activator family and are critical for cold acclimation due to binding a critical motif element in many cold-responsive genes. This element is called the C-repeat (CRT)/dehydration-responsive element (DRE) and contains the core sequence CCGAC (Baker et al. (1994); Yamaguchi-Shinozaki and Shinozaki (1994)). In addition to being important for cold acclimation, this motif is critical in plant dehydration response due to a strong overlap between those respective pathways (Yamaguchi-Shinozaki and Shinozaki (1994)).

A secondary pathway initiated during cold stress and acclimation is the ABA signalling pathway (Laang and Palva (1992); Xiong et al. (1999)). Previous research found ABA levels increased in response to cold with many cold regulated genes responding to ABA input (Lang et al. (1994); Gilmour and Thomashow (1991); Laang and Palva (1992); Gilmour et al. (1998)). In fact, there is an enrichment in the ABA binding element, ACGTGG/T, within many cold regulated genes in addition to the CRT/DRE element (Hannah et al. (2005)). As a result, it appears cold stress and acclimation is controlled, at least partly, through an ABA-dependent pathway. For a complete understanding of cold acclimation, it will be necessary to understand these two pathways and how they both affect cold gene expression.

1.6.2 Chromatin modifications and the cold acclimation response

Chromatin remodelling and histone modifications have been found to be extensively involved in stress and acclimation responses, including cold response (Lee et al. (2005); Kwak et al. (2007); Hu et al. (2011); Kim et al. (2004); Jung et al. (2013); Zhu et al. (2008)). Of the cold upregulated genes identified, many are involved in histone modifications and chromatin remod-

elling processes (Lee et al. (2005)). For instance, high mobility group (HMG) proteins, which are linked to extensive chromatin remodelling, are upregulated in response to cold within Arabidopsis (Kwak et al. (2007)). In addition, one cold-responsive gene (*DREB*) became accessible after cold-induced methylation and histone acetylation in its promoter (Hu et al. (2011)). Likewise, a cold-responsive gene, *FLOWERING LOCUS C* (*FLC*), which is heavily studied in flowering control was found linked to histone modifications (Kim et al. (2004)). In this case, the protein HOS1 interacts with a protein called FVE preventing the histone deacetylase, HDA6, from remodelling the chromatin at the *FLC* locus (Jung et al. (2013)). Lastly, a protein known as HOS15 is implicated in chromatin remodelling as a result of cold stress. In HOS15 mutants researchers observed a genome wide increase in histone 4 acetylation levels which localize with DHSs (Zhu et al. (2008); Rando (2007); Bell et al. (2011)). In summary, the literature documents a significant change in histone modifications and chromatin remodelling in response to cold stress and acclimation. In order to more completely understand plant cold acclimation transcriptional control, details of how chromatin modifications and chromatin remodelling occur, and how they subsequently effect DNA accessibility, will be necessary.

1.7 Research objective

This thesis set out to characterize and identify DHSs across the Arabidopsis root epidermis and endodermis under control and cold acclimation conditions while integrating generated data with supplemental transcriptomic and epigenetic data. The first objective of this study is to generate an analysis pipeline and program to utilize and analyze the new data format generated through the DNase-DTS protocol. The second objective, is to use the analysis pipeline to identify DHSs within the Arabidopsis root epidermis and endodermis under control and cold acclimation. The third objective, is to analyze and integrate the resulting data with public transcriptomic and epigenetic data to identify distinct associations and patterns. Through previous research indicating distinct differences in chromatin accessibility across tissue and cell-types

under various conditions, it is hypothesized there will be distinct differences in chromatin structure across the root epidermis and endodermis under cold and control conditions. In addition, it is expected each cell-type and condition will show many shared but many unique DHSs across the genome and that they will display unique properties. Through transcriptomic data integration, an association of DHSs with gene expression is expected to be observed. Specifically, highly transcribed genes are expected to be associated with the majority of highly accessible DHSs in their promoters. Likewise, through epigenetic data integration, an association of DHSs with histone modifications and DNA methylation is expected to be observed. The last objective of this study is to identify TF binding motifs enriched within accessible promoters of cell-type specific and cold acclimation specific genes. It is hypothesized these cell-type and acclimation specific genes will be associated with distinct TF binding motif patterns. Integration of DHS data and motif data with the cold acclimation CBF pathway will hopefully enable a greater understanding as to how this pathway is controlled through chromatin accessibility and TF binding.

The resulting *DNase-DTS* (*DDTS*) analysis pipeline and program resulting from this thesis will enable future studies into cell-type specific DNase I studies and be used with various data sources like ChIP-seq. The resulting DHS data from the epidermis and endodermis may be used as a resource for future studies looking into cell-type and stress responsive chromatin dynamics. Results from this study will enable a further understanding into the characteristics of transcription and epigenetics across cell-types and environmental conditions. By integrating epigenetic and transcriptomic data it will enable a higher resolution snapshot into distinct regulation patterns and the mechanisms of how they interact and are controlled. Lastly, this study outlines a possible way to utilize DHSs with motif mapping to select motif targets for transgenic applications.

Chapter 2

Materials and Methods

2.1 Plant growth conditions

T3 transgenic Arabidopsis seeds (Col-0 ecotype) were sterilized using chlorine gas in a gas chamber for fifteen minutes. Transgenic Arabidopsis seeds were imbibed in test tubes containing $\frac{1}{2}$ X MS salts for three days at 4°C. Seeds were sown on mesh squares approximately 1 inch by 1 inch and placed on agar plates containing $\frac{1}{2}$ X MS salts (Murashige and Skoog (1962)) with 50 µg/mL of kanamycin. Agar plates were allowed to germinate (in a growth chamber) at 24°C with 24 hours of light. Ten days after germination the mesh squares were transferred to 125 mL flasks with $\frac{1}{2}$ X MS salts and 1% sucrose. Silicosen® C-type caps were used to seal flasks in order to prevent contamination but enable ventilation. Flasks were shaken at 80 rpm at 24°C with long day conditions (18 hours of light, 6 hours of dark). Media was thereafter refreshed every three days. Control plants were grown in flasks for four weeks and cell-type specific nuclei were isolated using the INTACT protocol (see Section 2.2.2; Deal and Henikoff (2011)). In contrast, cold acclimated plants were grown for three weeks at room temperature and then grown at 4°C for one additional week. Once the week of acclimation was completed, cell-type specific nuclei were isolated.

2.2 DNase direct to sequencing

2.2.1 Transgenic Arabidopsis lines

Transgenic Arabidopsis cell lines for isolating cell-type specific nuclei were created using the floral-dip method (Zhang et al. (2006)). Transgenic Arabidopsis (ecotype Col-0) were transformed using a modified pCAMBIA 2300 plasmid containing a transgene designed for isolating tagged nuclei. The transgene encodes a nuclear envelope targeting protein composed of a nuclear envelope targeting signal, green fluorescent protein (GFP), and a biotin ligase recognition peptide (BLRP) (Deal and Henikoff (2011)). To enable cell-type specific isolation the transgene was transformed under the control of the *WEREWOLF* (AT5G14750) gene promoter for root epidermis expression, and the *SCARECROW* (AT3G54220) gene promoter for root endodermis expression (Lee and Schiefelbein (1999); Di Laurenzio et al. (1996)). Each transgene was co-transformed with a biotin ligase (BirA) from *E. coli*. All experiments were performed using T3 or higher transgenic lines. As a result, the cell-types of interest now express a BLRP in the nuclear membrane that acquires a biotin signal from the biotin ligase. Cell lines were confirmed through visual assessment of GFP expression under a Nikon® fluorescence microscope, model Eclipse Ni-U.

2.2.2 Isolation of cell-type specific nuclei

Isolation of cell-type specific nuclei was performed using the INTACT protocol of Deal and Henikoff (2011) with several changes. Root balls of approximately 4 g were cut from 6-12 plants grown in liquid media using a razor blade and ground in liquid nitrogen to a fine powder using a mortar and pestle. Ground tissue was transferred to another mortar and pestle and suspended in 10 mL of ice-cold nuclei purification buffer (NPB). NPB contains 20 mM of MOPS, 40 mM of NaCl, 90 mM of KCl, 2 mM of EGTA, and 0.5 mM of EDTA. MOPS was added to 250 mL of laboratory grade water and pH was adjusted to 7. Everything else was then added and final volume was brought up to 500 mL. NPB was filter sterilized through a

VWR®Bottle Top Filtration (500 mL, .2 μm pore size). Prior to performing INTACT, 0.2 mM of spermine, 0.5 mM of spermidine, and 1x proteins inhibitor cocktail are added to 40 mL of NPB.

The suspended tissue is now filtered through 70 μm and 40 μm filters and centrifuged at 1000 g for 10 minutes at 4°()C. Meanwhile 25 μL of streptavidin-coated beads (M-280, Invitrogen) were added to 1 mL of NPB and 10 μL of DAPI were added to 1 mL of NPB. Discard the supernatant from the tube in the centrifuge and re-suspend tissue pellet in 1 mL of NPB with DAPI, incubate for 3 minutes on ice. Centrifuge NPB with beads at 1000 g for 2 minutes at 4°C and discard supernatant, re-suspend in 25 μL of NPB. Centrifuge tissue with NPB and DAPI at 1000 g for 10 minutes at 4°C, discard supernatant. Suspend tissue in 1 mL of NPB and add beads. Rotate at 4°C using a rotating mixer for 45 minutes during which 0.1% (vol/vol) Triton X-100 is added to the remaining NPB.

Add suspended tissue to a 15 mL conical containing 9 mL of NPB and 0.1% (vol/vol) Triton X-100. Incubate tube for 10 minutes at 4°C using a using a DynaMag 15 (Life Technologies). Remove supernatant being careful not to disturb nuclei and add 10 mL of NPB with 0.1% (vol/vol) Triton X-100. Incubate again for 10 minutes at 4°C and remove supernatant. Suspend beads along tube walls with 500 μL of nuclease free water. Bead-Bound nuclei and unbound nuclei were counted using a Bright-Line® hemacytometer on a Nikon® fluorescence microscope, model Eclipse Ni-U. Purity of nuclei preparations were calculated by dividing the bead-bound nuclei by the total nuclei number.

2.2.3 DNase I digestion and isolation

Nuclei from each biological replicate were divided into four separate samples of 100,000 nuclei each. To identify an appropriate amount of digestion for each biological replicate in downstream bioinformatics, three samples were digested at varying amounts (0.1 U - 0.7 U) of DNase I in addition to an undigested sample (0 U). Nuclei were digested for 10 min at 37°C. After DNase I digestion, DNase I was inactivated using 50 mM EDTA and an incubation pe-

riod of 10 minutes at 70°C. DNA was isolated from nuclei using the QIAamp® DNA Micro Kit (Qiagen).

2.2.4 DNA sequencing using Nextera® and the Illumina® MiSeq®

Three digested samples and one undigested sample from the same biological replicate were prepared for sequencing on a single Illumina v3 600 sequencing cartridge using the 'Nextera® XT Library Prep Kit' (Illumina). Samples were identified by labelling them with separate indexes (Nextera Index Kit). The Agilent 2100 Bioanalyzer was used to analyze samples to ensure efficient DNase I digestion and sequencing library prep. A proper digestion profile is seen when there is a large peak of small sized fragments that slowly decrease towards larger fragments. An undigested sample is observed when there is a large peak of large sized fragments. It is important that a shift is observed from undigested to optimally digested across a range of DNase I units so an accurate assessment of digestion may take place. The Bioanalyzer analysis is only a first line check of digestion and a more comprehensive assessment occurs during computational analysis. DNase I analyses were repeated against fresh nuclei at higher units of DNase I if samples appeared poorly digested. After Nextera library preparation, the four biological-related samples were paired-end sequenced using 500 cycles of a MiSeq® Reagent Kit v3 600 cartridge on an Illumina MiSeq® sequencer. Digested samples from the same biological replicate were sequenced on the same sequencing cartridge to minimize sequencing biases.

2.2.5 Mapping sequencing reads to the reference

Raw reads produced by the Illumina MiSeq were retrieved in FastQ format. The raw paired-end reads were aligned to the TAIR10 (Berardini et al. (2015)) genome using BWA (Li and Durbin (2009)) with the following parameters:

```
-q 30 -t 15 -n 0.04 -o 1 -e -1
```

BWA was set to trim a read down to retain a Phred quality score >30. A phred score

above 30 retains only high quality base pairs with a probability of an incorrect base call of 1 in 1000 (Ewing et al. (1998)). PCR duplicates generated by PCR amplification were removed by identifying any read pairs with the same external coordinates. The pair of reads with the highest mapping quality were retained for further analysis. In addition, reads mapping to multiple genomic locations were removed so only uniquely mapped reads remained. Lastly, sample read counts were downsampled to the same read count for accurate sample to sample comparisons.

2.2.6 Computational analysis

Mapped reads were run through a custom script, called *DDTS*, which processes the raw data and identifies DHSs. *DDTS* was written in Python specifically for this project as no analysis tool has been developed for this form of data. The script has been packaged to be published and used as a Linux tool. Pseudo-code for *DDTS* is available in Algorithm 2.1. A detailed description of *DDTS* is available in the results section.

Briefly, *DDTS* identifies DHSs by statistically comparing digested samples to undigested samples and identifying regions within the digested sample with a significant lack of sequencing. DHSs are further filtered based upon a statistic, the likelihood ratio, comparable to a fold change in microarray or RNA-seq studies. *DDTS* outputs the genomic locations of DHSs and statistical information of each replicate in a BED file. Correlation between replicates is performed on processed wig format files to ensure high reproducibility. Replicates were run through *DDTS* on an individual basis to assess DNase I digestion and reproducibility of each replicate. To identify statistically significant DHSs and produce final DHS datasets, three replicates were run through *DDTS* in tandem. In combination to individual replicate assessments, optimal DNase I digestion profiles and sample reproducibility were assessed in final DHS datasets. Lastly, datasets were checked to ensure they lacked a DHS in the promoter of a negative control gene called *CINFUL-LIKE*, a transposable element (AT4G03770; Zhu et al. (2015); Shu et al. (2013)). As a transposable element, *CINFUL-LIKE* is a silent gene and was previously found associated with closed chromatin (Shu et al. (2013)). Likewise, datasets

were checked to ensure they contained a DHS in the promoter of a positive control gene called *ACTIN7* (AT5G09810; Zhu et al. (2015); Shu et al. (2013)). As *ACTIN7* is a gene expressed constitutively, it was previously found associated with open chromatin (Shu et al. (2013)).

2.2.7 DNase hypersensitive site analysis

To assure each DHS had a unique genomic location identifier, DHSs were classified into genomic categories in order of importance from the upstream 1000 bp, 5'UTR, 3'UTR, exon, intron, downstream 200 bp, and intergenic. Gene ontology (GO) analysis was performed using a custom script which tests for gene enrichment within GO slim categories using a hypergeometric test.

Statistical comparisons between sizes of DHSs across genomic categories were performed in R through a Kruskal-Wallis & Dunn test with a p-value <0.05. Correction of experiment-wise error-rate was performed using a Benjamini-Hochberg adjustment. The FSA and dunn.test R packages were used for statistical analysis of DHS size (Dinno (2017); Ogle (2017)).

input : Undigested .bam and digested .bam files or undigested and digested .wig files
output: .wig file when input is .bam or .bed file when input is .wig

```
if input is .bam then
    Convert input to .wig using F-Seq
end
if input is .wig then
    Read in 100,000 bp basepairs of sequence for all WIG files and normalize start
      position across replicates
    while file end is not reached do
        foreach Scanning window do
            foreach Base pair in window do
                foreach Biological replicate do
                    Append kernel probability of current base pair for each undigested
                      and digested replicate to storage array
                end
            end
        end
        Take mean of current scanning window of each undigested and digested replicate
        if Number of replicates >= 3 then
            Perform T-test between undigested and digested sample means
        else
            else we perform a T-test between the scanning windows of digested and
              undigested samples
        end
        if Test Statistic > cutoff then
            flag = TRUE
            foreach Biological Replicate do
                if LikelihoodRatio < cutoff then
                    flag = FALSE;
                end
            end
            if flag = TRUE then
                Append DHS location, test statistic, p-value, and mean replicate kernel
                  probabilities to the output file
            end
        end
        Read in next 100,000 bp chunk from the .wig files. Repeat loop until end of .wig
          files reached
    end
    Return BED file with DHS position and statistical data
end
```

Algorithm 2.1: Main Function of DNase-DTS

2.3 Transcriptomics

2.3.1 Processing RNA-seq data

Raw RNA-seq data were obtained from Li et al. (2016). Briefly, this study obtained RNA from 15 distinct root cell-types using fluorescence-activated cell sorting combined with paired-end sequencing their samples on an Illumina HiSeq2000. For this work, raw paired reads were obtained from the NCBI SRA database under BioProject PRJNA323955 and mapped to the TAIR 10 genome using *STAR* (Dobin et al. (2013)) using the following parameters:

```
STAR --genomeDir star-genome --outFilterMultimapNmax
20 --alignSJoverhangMin 8 --alignSJDBoverhangMin8
--outFilterMismatchNmax 8 --alignIntronMin 35 --alignIntronMax
2,000 --alignMatesGapMax 100,000 --readFilesCommand zcat
--runThreadN 6 --readFilesIn
```

For a detailed description of what each parameter entails see Dobin et al. (2013) and Li et al. (2016). Only uniquely paired reads were retained for further analysis. For each gene, the number of mapped reads were counted through the featureCounts script which proceeds to output the number of uniquely mapped reads for each gene (Liao et al. (2013)). Finally, the read count for each gene is used to calculate the FPKM (fragments per kilobase of transcript per million mapped reads) with the following equation:

$$FPKM_i = \frac{X_i}{\tilde{l}_i N} * 10^9 \tag{2.1}$$

Where i is the current transcript, X_i are the read counts for each transcript, $\tilde{l}_i$ is the effective length, and N is the number of reads. To control for biological variation the mean FPKM of three replicates is used for all analyses. The FPKM is used for this expression analysis as it enables normalization between samples and accurate comparison between genes. To integrate DHS data with gene expression data, genes were sorted from highest to lowest FPKM and

split into six separate expression bins (see Figure 3.7). An identical analysis was performed on epidermal and endodermal microarray data to confirm the results with a secondary source. Processed epidermal and endodermal microarray data were obtained from Birnbaum et al. (2003).

Despite cold RNA-seq data not existing for cell-type specific nuclei, the prior analysis was repeated by integrating cold DHS data with cold acclimated Arabidopsis microarray data obtained from Hannah et al. (2005). Raw microarray data were loaded into R (Team (2013)) and analyzed using the R packages *affy* (Gautier et al. (2004)) and *ath1121501.db* (Carlson (2016)). For cold microarray data, the mean probe intensity was calculated from the mean of three replicates. Similar to the RNA-seq data, genes were sorted from highest to lowest expression level and associated with respective DHSs (see Figure 3.8).

ANOVA analysis was performed to identify if DHS t-test scores or likelihood ratios were associated with gene expression levels. A likelihood ratio was chosen as an indirect measure of a DHS's accessibility and is obtained by dividing the mean of a DHS's control kernel probability by the mean of the DHS's digested kernel probability (see Results). To accomplish this, DHSs were associated with their respective gene and FPKM, sorted from highest to lowest t-score or likelihood ratio, and split into six groups or bins (see Figure 3.9). To normalize deviations, FPKM for all genes were $\log_{10}$ transformed. Deviations within this analysis were assumed to have constant variance, be independent, and normally distributed with a mean of zero. Normality assumptions were tested through visual analysis of the residuals against the fitted values in addition to a Q-Q plot. To test for the assumption of constant variance a Bartlett's test was performed. Mean FPKM expression values for each treatment and 95% confidence intervals were calculated using the R package *lsmeans* (Lenth (2016)). Fitted means on the log scale were back-transformed to their natural scale before visual plotting.

2.3.2 Identifying differentially-expressed genes

The R package DESeq was used to identify differentially-expressed root epidermal and endodermal genes from RNA-seq data obtained from Li et al. (2016) (Anders and Huber (2010)). DESeq identifies differentially-expressed genes using raw count data and a negative binomial distribution. Differentially expressed genes were identified using a binomial test with adjusted p-values (p <0.01) using a Benjamini-Hochberg adjustment. Cold differentially-expressed genes were obtained from Hannah et al. (2005).

2.4 Epigenetics

Processed and raw Methyl-seq data were obtained from Kawakatsu et al. (2016) through the NCBI GEO database under GSE79710. Processed data contained only identified methylated cytosines and was used for integration with DHSs. Methylated cytosines were identified through a binomial distribution as described in Kawakatsu et al. (2016). Cytosines were labelled as identified methylated cytosines if they passed a significance threshold of $p < 0.01$.

For comparing DHSs with all mapped cytosine positions, raw data from Kawakatsu et al. (2016) was processed and analyzed. The *bismark* program utilizing the *bowtie2* script mapped the raw Methyl-seq fastq data to the Arabidopsis TAIR10 genome to obtain a site by site methylation profile (Langmead and Salzberg (2012); Krueger and Andrews (2011)) through the following command:

```
bismark --bowtie2 -n 1 -l 50
```

Running bismark this way does not allow any more than one non-bisulfite mismatch for every read. However, this alone does not result in the basepair resolution methylation data. To accomplish site specific CpG, CHH, and CHG methylation data, the processed data were filtered through the *bismark_methylation_extractor* script with the following parameters:

```
bismark_methylation_extractor --multicore 8 --comprehensive
--bedgraph --counts --cytosine_report --CX_context
```

Methylation information for each cytosine was only retained if the specific cytosine had at least four mapped reads.

Processed histone data were obtained from Deal and Henikoff (2010) under the GSE accession number GSE19654. Briefly, this experiment used the INTACT protocol to isolate epidermal non-hair cell nuclei using the GL2 promoter. DNA was then isolated from nuclei using ChIP with antibodies specific to the required histone modification. Isolated DNA was cohybridized to a custom Roche NimbleGen Arabidopsis tiling array with H3 ChIP DNA. This thesis took the mean of two biological replicates for each ChIP experiment and integrated it with DHS data.

2.5 TF binding site enrichment

A-priori motif prediction was performed on 750 bp upstream and 250 bp downstream of gene transcriptional start sites (TSSs) using *Cismer* and *pssmROC* (Austin et al. (2016)). *A priori* motifs tested for enrichment were provided by Weirauch et al. (2014). This motif list contains hundreds of motifs across 32 transcription factor families. Motifs were tested on an individual basis for enrichment and enriched motifs summarized using their respective transcription factor family for plotting (see Figure 3.17). The custom script, *pssmROC* was run on gene specific lists using the following parameters:

```
pssmROC weirauch.pssms background.fasta geneList.fasta 0 1
```

The weirauch.pssms is a file containing the positional specific scoring matrices (PSSM) for each motif in the dataset. The background.fasta file contains the upstream 750 bp and the downstream 250 bp of each gene TSS in the TAIR10 genome. The geneList.fasta file contains the upstream 750 bp and the downstream 250 bp of each gene TSS in a desired gene list. The 0 and 1 indicate the range of functional depths at which to test for motif enrichment. Motif locations in every gene were plotted in addition to DHS locations for comparative analysis. Motifs were identified as enriched if they were found to have a Z-score >3.

Chapter 3

Results

This work developed a custom script called *DDTS* for the efficient and accurate identification of DHSs across the Arabidopsis genome using sequencing data obtained from nanogram quantities of DNA isolated from cell-type specific nuclei preparations. *DDTS* statistically identified DHSs through a novel bioinformatic tool which incorporates a novel analysis algorithm, F-Seq (Boyle et al. (2008b)), and BEDtools (Quinlan and Hall (2010)). Computational quality checks, in addition to *DDTS*, were developed to ensure proper DNase I digestion. Results identified several thousand DHSs in the Arabidopsis root epidermis and endodermis under control and cold conditions. Analysis of DHSs revealed distinct DHS genomic distribution profiles with the majority of DHSs surrounding the TSS. A comparison of DHS size within genomic locations identified significantly wider DHSs within the upstream 1000 bp. Integrating DHSs with RNA-seq, microarray, methylation, and histone modification data resulted in correlations with gene expression and unique epigenetic patterns. Integration with RNA-seq and microarray data obtained differentially-expressed and simultaneously differentially-accessible (DE/DA) genes from each cell-type and condition. Testing for enrichment of TF binding motifs within DE/DA genes identified enrichment of motifs from the TF families AP2, bHLH, bZip, and CG-1.

3.1 Identifying DHSs in the Arabidopsis genome with *DDTS*

The *DDTS* computational analysis pipeline was developed to process data generated through the DNase-DTS protocol and identify DHSs. *DDTS* was written in Python and has been packaged to be published and used as a Linux tool in combination with existing Linux applications. It may be used on all organisms and across different experiments generating data requiring an identification of a lack of sequencing.

3.1.1 Raw data conversion and processing

DDTS's first step is to convert files containing uniquely mapped reads, in BAM format, to BED format using BEDTools *bamToBed* (Quinlan and Hall (2010)). Converted files are subsequently run through a program called F-Seq with the following parameters: -v -l 200 (Boyle et al. (2008b)). F-Seq converts raw mapped reads in BED format to a continuous base pair probability signal, in WIG file format. In other words, F-Seq calculates a Gaussian kernel density estimation which generates a continuous and smooth signal. This smooth signal is the probability for every base pair or, is the probability or likelihood of a sequenced read being found at that specific base pair. Wig files for each sample and separate chromosome are generated from this step. Correlation analysis between replicates, using the wig files, was performed to ensure replicates are highly reproducible.

3.1.2 Novel algorithm for identification of DHSs

After the processing of raw data, *DDTS* identifies DHSs by scanning and comparing the kernel probability of the digested and undigested samples. *DDTS* runs each chromosome separately on distinct CPU cores for quick and efficient analysis times. To keep memory usage low, for use on a wide range of computers, *DDTS* reads in 100,000 base pairs at a time from the wig files. *DDTS* scans along the genome in specified window sizes and shifts this window a small step at each iteration. For example, all analyses performed in this work were performed with a

scanning window size of 300 bp and a step size of 25 bp. As a result, these settings limit DHS size identification to a minimum of 300 bp. However, this window and step size may be altered depending on the initial data. At this stage *DDTS* reports DHSs through one of two methods depending on the number of replicates.

If there are less than three replicates, *DDTS* compares all probability values in the current 300 bp window of the digested sample to the 300 bp window of the undigested sample through a t-test. If three replicates or more are inputted, a mean of the 300 bp window for each replicate is taken. The t-test for this method is performed between means of the undigested replicates and means of the digested replicates. For both methods, if the t-test reports a t-score higher than the specified cut-off it continues to filter the data using a second filter.

The second filter is similar to the fold difference cut-off in RNA-seq data, called the likelihood fold difference or likelihood ratio. *DDTS* calculates the likelihood ratio for a DHS by dividing the mean probability of finding a read in the undigested samples by the probability of finding a read in the digested samples. To ensure a stringent identification procedure, each replicate has to pass this likelihood ratio cut-off individually. DNase-DTS compares the undigested and digested sample from the first replicate, checks to see if it passes the cut-off, and continues on to the other replicates. If all replicates pass this cut-off, a DHS is reported for the current window.

To control the false discovery rate (FDR) of the analysis, the likelihood ratio is adjusted until a sufficient FDR is met. In general, the higher the ratio cut-off the lower the false discovery rate. However, increasing the cut-off results in a lower true positive rate thereby trading sensitivity for increased specificity. As a result, to keep the true positive rate as high as possible, a cut-off enabling a 5% false discovery rate (FDR <.05) was selected. In addition, the likelihood ratio is proportional to a DHS's accessibility and will be discussed further later. For each window, the statistical test and filters are performed until the current loaded 100,000 bp sequence is read, after which, the next 100,000 bp sequence is read in. This analysis is repeated for each 100,000 bp sequence until the entire length of the chromosome is analysed.

After the entire chromosome has been read, *DDTS* outputs a file of identified DHSs. *DDTS* reports the location of each DHS, the t-score, p-value, and mean probability for each undigested and digested replicate. Since the step size results in overlapping windows the reported DHSs may overlap. As a result, BEDTools *mergeBed* combines the overlapping windows post-analysis. Provided in the methods section is the pseudo-code describing the main function or algorithm of the *DDTS* script when run with input WIG files, helper functions for reading and writing files are left out for clarity (Algorithm: 2.1).

3.2 Assessing optimal DNase I digestion and computational settings

Proper library prep and DNase I digestion of biological samples were analyzed on an Agilent® Bioanalyzer. Observations of samples exposed to a DNase I dilution series revealed a shift from a higher frequency of larger DNA fragments to a higher frequency of smaller DNA fragments (Figure 3.1). The appropriate digestion profile is evident when one observes a large peak of small sized fragments that slowly decreases as you get to larger fragments. However, over digested samples will also display a large peak of small sized fragments. Hence, this assessment should only be used in a dilution series where one may observe the shift across a range of DNase I units. Figure 3.1 shows an optimally 0.5 U digested sample, as it displays a high frequency of small fragments. Figure 3.1 also displays a decreased amount of DNA across the dilution series, indicated by a decrease in fluorescence units. The Bioanalyzer step is optional as ultimate assessment of DNase I digestion takes place through computational analysis.

After ensuring proper library prep, digested DNA samples were sequenced on an Illumina® MiSeq NGS system. Table 3.1 shows a brief summary of sequencing information of each biological replicate and experimental condition across the DNase I dilution series. Sequencing obtained 676 million reads from all sequencing libraries. From these 635 million reads (94%) mapped to the Arabidopsis genome with 480 million reads (71%) mapping to a unique se-

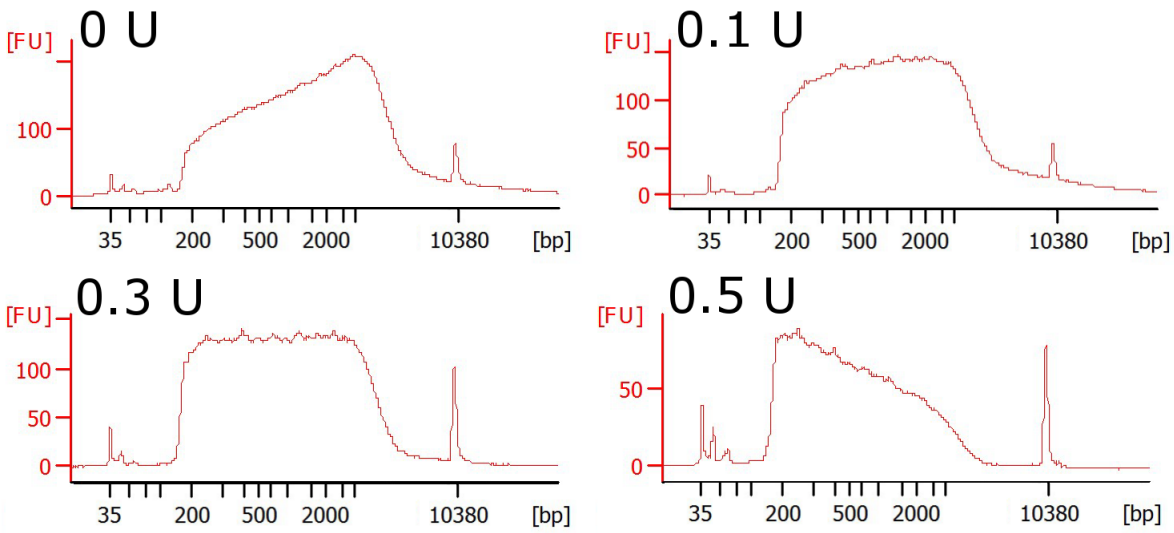


Figure 3.1: **Bioanalyzer results for one biological replicate digested at 0.0 U, 0.1 U, 0.3 U, and 0.5 U of DNase I.** The fragment migration time (seconds) versus the fluorescence units is shown. The x-axis represents the size of DNA fragments while the y-axis represents the quantity of fragments. Peaks observed on either end of each subplot are label markers. Increased digestion shifts fragment size distribution to a higher frequency of small sized fragments.

quence. One outlier to note is the endodermal cold replicate #3 had a low unique read percentage when compared with other sequencing runs. As the 0.5 U sample resulted in a similar read count to other runs, the data from this sample was used in downstream analysis (see Figure 3.2).

The *DDTS* program processed and analyzed DHSs of individual replicates to assess their digestion profiles across a DNase I dilution series (Table 3.2). Replicates showed a typical digestion pattern of an increasing number of DHSs as DNase I units increased. In addition, each replicate displayed an increasing width of DHSs as DNase I units increased. These results indicate an increase in the amount of digested DNA due to higher units of DNase I. Results identified a greater number of DHSs and larger DHSs in endodermal cold replicate #3, an outlier compared to other replicates. It is important to note the endodermal cold biological replicates were digested with a different amount of DNase I compared to other samples at 0.3 U, 0.5 U, and 0.7 U of DNase I. Before continuing the analysis, individual replicates were assessed for optimal DNase I digestion through computational analysis. Appendix A displays the complete digestion profiles of each replicate in pie and bar chart figures. Replicates showed ideal DNase I digestion profiles across DNase I dilution series. Each replicate displayed an increase in percent of DHSs within the upstream 1000 bp and a decrease in percent of DHSs within other genomic regions. These figures may also be used to assess the number and size of DHSs in a dataset, as was done in Table 3.2.

Individual replicates were correlated using processed bigWig format files to ensure high reproducibility before continuing the analysis with combined replicates. Table 3.3 shows the correlation between each replicate in each cell-type and experimental condition. All replicates demonstrated high correlation with one outlier, the endodermal cold replicate showed lowest correlation with r=0.829 at 0.7 U of DNase I. To additionally assess high reproducibility, replicates were compared using all available data including DHS distribution, size, and number of DHSs.

Running individual replicates through *DDTS* is used to assess reproducibility and DNase I digestion of individual replicates. To identify optimal DHS data for each cell-type and ex-

Table 3.1: **Summary of sequencing data from each biological replicate and experimental condition.**

	Total Reads[a]				Mappable Reads[b]								Unique Reads[c]							
DNase I Units	**0 U**	**0.1 U**	**0.3 U**	**0.5 U**	**Total**	**0 U**		**0.1 U**		**0.3 U**		**0.5 U**		**0 U**		**0.1 U**		**0.3 U**		**0.5 U**	
Epidermal control replicate #1	13990358	14221268	10854262	9845388	48911276	13325927	95.25%[d]	13679528	96.19%	10335188	95.22%	9057774	92.00%	10212047	72.99%	11558270	81.27%	8642423	79.62%	6260538	63.59%
Epidermal control replicate #2	12971146	10314816	10391222	11799498	45476682	12262182	94.53%	9860149	95.59%	9914711	95.41%	11134029	94.36%	9534018	73.50%	8177209	79.28%	8291969	79.80%	8847233	74.98%
Epidermal control replicate #3	13918512	13754854	14909176	13178532	55761074	13219371	94.98%	13056117	94.92%	14151843	94.92%	12339600	93.63%	10483418	75.32%	10371178	75.40%	11313481	75.88%	9001012	68.30%
Endodermal control replicate #1	13483300	12942778	12775614	13261548	52463240	12765293	94.67%	12120224	93.64%	12240483	95.81%	12477367	94.09%	9791541	72.62%	9714467	75.06%	10322482	80.80%	9806392	73.95%
Endodermal control replicate #2	22442314	891020	12272844	14530644	50136822	21300490	94.91%	853263	95.76%	11825217	96.35%	13763429	94.72%	17015751	75.82%	698529	78.40%	10168767	82.86%	11009543	75.77%
Endodermal control replicate #3	13547706	13188880	12430546	15678664	54845796	12833111	94.73%	12584710	95.42%	11917522	95.87%	14805124	94.43%	10192324	75.23%	10500011	79.61%	10269971	82.62%	11793220	75.22%
Epidermal cold replicate #1	14082700	18302988	30774760	31753434	94913882	13369750	94.94%	16932352	92.51%	28047193	91.14%	28568054	89.97%	10407077	73.90%	12456040	68.05%	16250170	52.80%	14058019	44.27%
Epidermal cold replicate #2	25839020	13942452	6853768	10049864	56685104	24507701	94.85%	13287214	95.30%	6501466	94.86%	9488237	94.41%	18767898	72.63%	10685439	76.64%	5387908	78.61%	7786853	77.48%
Epidermal cold replicate #3	15560416	16373802	12272530	13913036	58119784	14552820	93.52%	15365525	93.84%	11716186	95.47%	12948552	93.07%	11491104	73.85%	12879388	78.66%	10098324	82.28%	10023721	72.05%
DNase I Units	**0 U**	**0.3 U**	**0.5 U**	**0.7 U**	**Total**	**0 U**		**0.3 U**		**0.5 U**		**0.7 U**		**0 U**		**0.3 U**		**0.5 U**		**0.7 U**	
Endodermal cold replicate #1	13502712	14947244	13097154	12298846	53845956	12744416	94.38%	14372662	96.16%	12356007	94.34%	11348002	92.27%	10406973	77.07%	12620430	84.43%	9857200	75.26%	8131806	66.12%
Endodermal cold replicate #2	13454004	12831018	9681196	11536026	47502244	12756074	94.81%	12287975	95.77%	9175323	94.77%	10784996	93.49%	10371963	77.09%	10722339	83.57%	7501300	77.48%	8213676	71.20%
Endodermal cold replicate #3	12980626	17165790	17052248	10463772	57662436	12227033	94.19%	15638259	91.10%	15561689	91.26%	9226574	88.18%	9646553	74.32%	9823846	57.23%	6867856	40.28%	2042818	19.52%

[a] The total number of reads sequenced.

[b] Reads mapped to the TAIR10 genome.

[c] Reads matching one location in the TAIR10 genome .

[d] Percentages shown are the percent mapped/unique reads compared to the total reads obtained.

Table 3.2: **Summary of the number of DHSs and their mean size, in bp, from each biological replicate and experimental condition.**

DNase I Units	**0.1 U**		**0.3 U**		**0.5 U**	
Epidermal Control	**Number of DHSs**	**Mean Site Size (bp)**	**Number of DHSs**	**Mean Site Size (bp)**	**Number of DHSs**	**Mean Site Size (bp)**
Replicate #1	32559	502.92	35488	502.43	54379	637.36
Replicate #2	30708	478.52	34369	519.33	35585	551.21
Replicate #3	27429	453.61	33030	491.67	44078	591.91
Endodermal Control						
Replicate #1	23889	446.26	25951	521.76	31233	536.12
Replicate #2	-	-	25690	564.91	31581	605.21
Replicate #3	23772	458.81	26486	525.52	33124	612.47
Epidermal Cold						
Replicate #1	33928	490.92	48909	747.68	50363	991.38
Replicate #2	28494	454.37	31085	471.49	37558	549.10
Replicate #3	28692	500.39	29792	551.14	37698	550.15
Endodermal Cold[a]	**0.3 U**		**0.5 U**		**0.7 U**	
Replicate #1	33432	568.02	38543	601.44	42209	801.92
Replicate #2	36962	558.27	37373	506.07	40428	596.60
Replicate #3	48096	1127.13	58178	1294.37	49145	1934.76

Note: All replicates were analyzed through *DDTS* with a t-score cut-off of 15 and a likelihood fold difference of 2.0.

[a] The endodermal cold replicates were run at 0.3 U, 0.5 U, and 0.7 U of DNase I.

Table 3.3: **Correlation analyses between biological replicates for each experimental condition.**

	Replicate Combination	Epidermal Control	Endodermal Control	Epidermal Cold Acclimated	Endodermal Cold Acclimated[a]
0 U of DNase I	**#1 vs #2**	0.981	0.960	0.959	0.980
	#1 vs #3	0.975	0.978	0.974	0.967
	#2 vs #3	0.982	0.969	0.964	0.966
0.1 U of DNase I	**#1 vs #2**	0.960	0.948	0.969	0.875
	#1 vs #3	0.964	0.967	0.948	0.858
	#2 vs #3	0.975	0.947	0.956	0.839
0.3 U of DNase I	**#1 vs #2**	0.964	0.921	0.944	0.950
	#1 vs #3	0.957	0.944	0.894	0.888
	#2 vs #3	0.961	0.923	0.938	0.846
0.5 U of DNase I	**#1 vs #2**	0.960	0.964	0.9351	0.973
	#1 vs #3	0.978	0.972	0.955	0.829
	#2 vs #3	0.971	0.963	0.956	0.829

Note: Correlation was performed on processed bigWig files for each replicate.

[a] Note the endodermal cold acclimated sample was digested at 0.3 U, 0.5 U, and 0.7 U units of DNase I.

perimental condition, replicates were run through *DDTS* in tandem across the DNase I dilution series. Final DHS data for each sample were chosen from the DNase I dilution series through analysis of DNase I digestion patterns. The optimal DHS data is the sample showing an optimal amount of DNase I digestion such that DNA is sufficiently digested but not over digested (see Methods). 0.5 U of DNase I was selected as the appropriate amount of digestion when assessing DHS datasets for 100,000 isolated nuclei (Figure 3.2). 0.7 U was found inappropriate for further analysis as it was found to be over digested. The discussion section discusses the reasoning behind selection of the optimal digested sample in greater detail. Selection of the optimal digested sample is important as it will enable a higher quality of identified DHSs. Over digestion results in regions not accessible being digested and identified as DHSs, while under digestion results in regions accessible not being digested and identified as DHSs. Thus, optimal digestion results in low false positives and low false negatives. Briefly, 0.5 U had the highest percentage of upstream 1000 bp DHSs and the smallest percentage of DHSs in other genomic categories. While 0.1 U and 0.3 U digested samples were not over digested, 0.5 U had a higher degree of digestion enabling efficient identification of DHSs. In addition 0.1 U and 0.3 U digested samples were under digested and identification of all DHSs would be difficult. 0.7 U had a drop in the percentage of upstream 1000 bp DHSs and a larger percentage of DHSs and was deemed too over digested to be used in the final analysis (see 'Epidermis and endodermis DHSs share distinct characteristics'). Repeating the dilution series for all biological samples was performed as it serves as an important quality control step in DNase digestion. It is important to select the appropriate digested sample for analysis as under digestion or over digestion will complicate DHS analysis and introduce high false negative rates.

To keep the false discovery rate (FDR) lower than 5%, the likelihood ratio cut-off in *DDTS* was adjusted and the number of DHSs identified in final datasets compared to the number of DHSs identified from equally sized random datasets (Table 3.4). Random sequencing datasets are randomly generated from existing data using BEDTools *random* and run through DNase-DTS exactly as performed to biological sequencing data (Quinlan and Hall (2010)). The t-

test statistically identified significant DHSs as long as the t-test reported a t-score above 3. Increasing the t-score only slightly altered the FDR compared with adjusting the likelihood ratio and so remained at 3 for all DHS datasets. Adjusting the likelihood ratio reduced the true positive rate (TPR), therefore, a balance was made between lowering the FDR and maximizing the TPR. The selected fold difference or likelihood ratio cut-offs for the epidermal control was 1.6, 1.5 for the endodermal control, 1.5 for the epidermal cold, and 1.6 for the endodermal cold. The final number of identified DHSs for each condition at 0.5 U of DNase I were: 18599 for the epidermal control, 16566 DHSs for the endodermal control, 16285 DHSs for the epidermal cold, and 15834 DHSs for the endodermal cold. These datasets at 0.5 U of DNase I were used for subsequent analyses.

Table 3.4: **Summary of the false discovery rate analysis for all experimental conditions at 0.5 U of DNase I.**

Experimental Condition	Fold Difference[a]	DHSs [b] (Biological Datasets)	DHSs[c] (Random Datasets)	FDR[d]
Root Epidermal Control	1	32074	22663	70.66%
	1.5	20621	1294	6.28%
	1.6	**18599**	**500**	**2.69%**
	1.7	16946	178	1.05%
	1.8	15552	75	0.48%
	2	13270	13	0.10%
Root Endodermal Control	1	27405	23311	85.06%
	1.4	18560	1665	8.97%
	1.5	**16566**	**508**	**3.07%**
	1.6	14891	164	1.10%
	1.7	13553	47	0.35%
	1.8	12286	12	0.10%
	2	10248	1	0.01%
Root Epidermal Cold	1	26401	23247	88.05%
	1.5	**16285**	**625**	**3.84%**
	1.6	14205	183	1.29%
	1.7	12571	57	0.45%
	1.8	11241	10	0.09%
	2	9166	1	0.01%
Root Endodermal Cold	1.0	25954	22023	84.85%
	1.5	17643	1499	8.50%
	1.6	**15834**	**637**	**4.02%**
	1.7	14152	251	1.77%
	1.9	11706	32	0.27%

Note: Selected cut-offs for final datasets are bolded. DHSs identified have a t-score >3.

[a] To lower the FDR of each sample below 5% the likelihood ratio cut-off was adjusted. The fold difference is the the ratio of the undigested probability value over the digested probability value.

[b] The number of DHSs identified for biological datasets.

[c] The number of DHSs identified for random datasets. Random datasets are generated with an identical number of reads to the biological samples.

[d] FDR is calculated as the number of DHSs identified in a random dataset divided by the number of DHSs in the biological samples.

3.3 Epidermis and endodermis DHSs share distinct characteristics

Final DHS data and DNase I digestion profiles for each cell-type and condition were analyzed to identify shared and unique characteristics. A summary of DNase I digestion profiles for each dataset are shown in Figure 3.2. Initial analysis of the bar plots in these figures revealed the number and size of DHSs increased as the units of DNase I increased, reconfirming the previous findings in Table 3.2. Additionally, the percentage of upstream 1000 bp DHSs increased as the DNase I units increased. As a consequence, the percentage of DHSs within other genomic locations decreased as the units of DNase I increased. However, Figure 3.2d revealed that when a large amount of DNase I is used, the sample can become over digested resulting in the percentage of upstream 1000 bp DHSs decreasing. The percentage of promoter DHSs increased within this sample except when digested with 0.7 U of DNase I. At 0.7 U of DNase, the percentage of promoter DHSs decreased by 20% and the percentage of exonic DHSs and intergenic DHSs increased.

Each cell-type and experimental condition displayed a shared genomic distribution of DHSs when exposed to identical DNase I units. For example, all samples digested with 0.5 U of DNase I contain 65%-70% of upstream 1000 bp DHSs (Figure 3.2). The epidermal and endodermal cell-types under control conditions display a similar distribution with all genomic locations within a couple percent of each other (Figure 3.2ab). The digestion pattern in Figure 3.2b at 0.3 U of DNase I displayed atypical results compared with other samples. Notably, a drop in the upstream 1000 bp DHSs were displayed at 0.3 U of DNase I. This atypical pattern was also observed in individual replicates of this sample. The reason for this result is unclear, but may indicate a biological phenomenon as it occurred in all three separate biological replicates.

The exact number of DHSs across genomic categories and the number of genes containing a DHS across genomic locations are shown in Table 3.5. In general, the number of DHSs de-

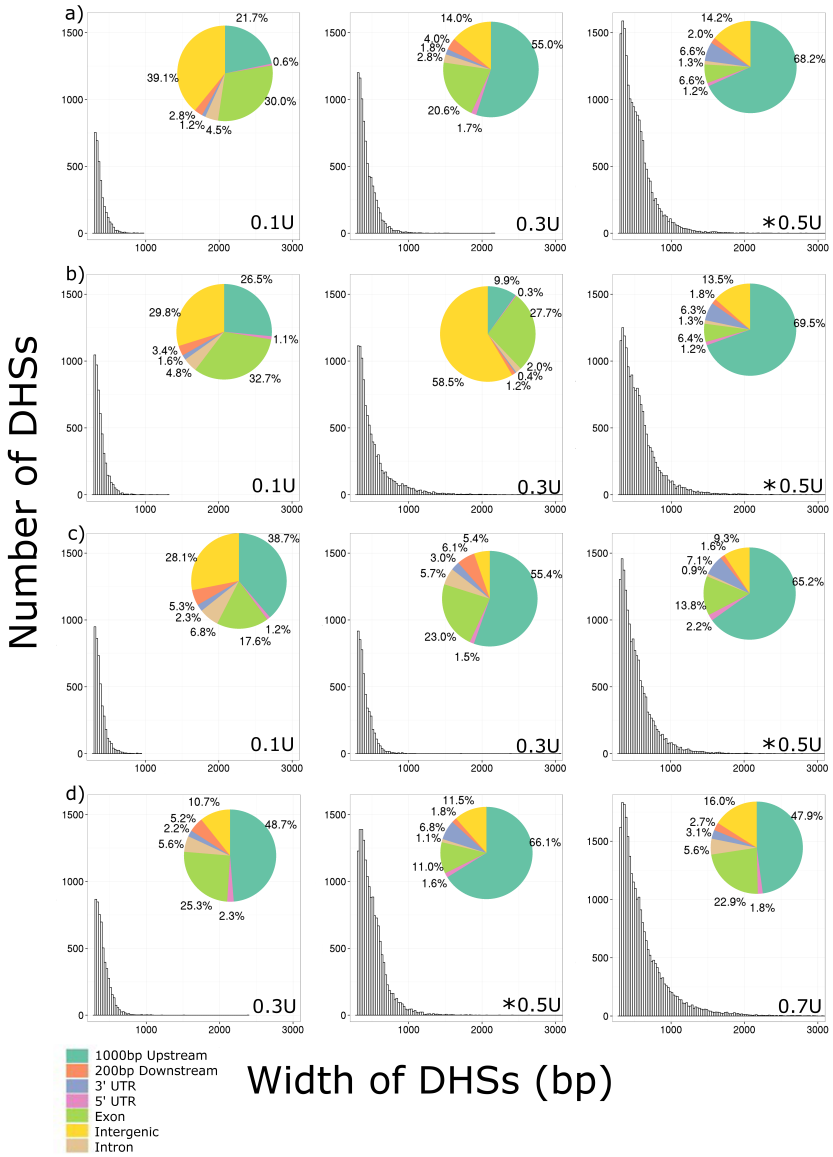


Figure 3.2: **DNase I digestion profiles of each cell-type and experimental condition.** DNase-DTS bar plots displaying the frequency of DHSs with respect to DHS size. Also displayed are pie graphs showing the distribution of DHSs within genomic locations. DHSs were classified into genomic categories in order of importance, upstream 1000 bp, 5'UTR, 3'UTR, exon, intron, downstream 200 bp, and intergenic. Concentration of DNase I shown in lower right of each subplot. Asterisk indicates chosen dataset for downstream analysis. a) Digestion profile from epidermis under control conditions. b) Digestion profile from endodermis under control conditions. c) Digestion profile from epidermis under cold conditions. d) Digestion profile from endodermis under cold conditions.

creased from control to cold acclimated datasets. Results revealed the number of exon DHSs and genes within the cold datasets are more numerous compared with the number of control exon DHSs and genes. Additionally, while there is an overall reduction in upstream DHSs between control and cold conditions, the number of genes associated with an upstream 1000 bp DHS was not altered significantly. 14,071 genes with an upstream 1000 bp DHS in epidermal control compared to 13,477 in epidermal cold conditions. Reconfirming the previous results, the endodermal control dataset contained 13,112 genes with an upstream 1000 bp DHS compared with 13,443 genes in the endodermal cold dataset.

DHS datasets from each experimental condition and cell-type were compared visually and statistically. Distributions of DHSs across each chromosome and the overlap between datasets are shown as Hilbert plots in Figure 3.3. Hilbert plots were built for all conditions and cell-types but were all similar and so only the epidermal and endodermal datasets are shown. Hilbert curves show each chromosome not as a typical linear representation but as a recursively built fractal. This enables a 2D visualization of each chromosome allowing one to infer spatial properties by keeping neighbouring positions on each chromosome within close proximity (Gu et al. (2016)). A distinct observation from these figures is the lack of DHSs within the centromeres, or condensed heterochromatin, of each chromosome. Comparisons between each dataset display a large amount of overlap but also show distinct differences between each cell-type and between cold and control conditions (Figure 3.3c). The epidermal and endodermal datasets significantly overlapped with 11,823 shared DHSs, 6,776 unique epidermal DHSs (36.4%), and 4,743 unique endodermal DHSs (28.6%) ($p < 0.001$, Fishers exact test). The epidermal control and cold datasets significantly overlapped with 10,112 DHSs shared between these two respective datasets, 8,487 (45.6%) unique epidermal control DHSs, and 6,173 (37.9%) unique cold epidermal DHSs ($p < 0.001$, Fishers exact test). In other words, 8,487 DHSs closed and 6,173 DHSs opened in the epidermis in response to cold. Repeating this with the endodermal control and cold dataset reveals significant overlap with 9,425 DHSs shared, 7,141 (43.1%) unique to the endodermal control dataset, and 6,409 (40.5%) unique to the endodermal cold

Table 3.5: **Number of DHSs within each cell-type and experimental condition and the associated number of genes across each genomic location.**

Genomic Location	Total	Upstream 1000 bp	5'UTR	3'UTR	Exon	Intron	Downstream 200 bp	Intergenic
Number of Epidermal DHSs	18599	12678	215	1232	1219	236	376	2643
Number of Epidermal Genes	16951	14071	232	1515	1229	222	393	-
Number of Endodermal DHSs	16566	11511	201	1049	1055	208	298	2244
Number of Endodermal Genes	15850	13112	217	1271	1042	202	317	-
Number of Epidermal Cold DHSs	16285	10610	352	1151	2249	152	263	1508
Number of Epidermal Cold Genes	15439	13477	385	1466	2185	152	280	-
Number of Endodermal Cold DHSs	15834	10474	251	1079	1748	179	278	1825
Number of Endodermal Cold Genes	15281	13443	283	1355	1741	176	294	-

Note: DHSs were classified into genomic categories in order of importance, upstream 1000 bp, 5'UTR, 3'UTR, exon, intron, downstream 200 bp, and intergenic. Due to intergenic DHSs not being associated to any gene, this information has been left out.

dataset ($p < 0.001$, Fishers exact test).

The distribution of DHSs around the TSS were investigated since, as expected, the majority of DHSs are located within gene promoters (Figure 3.2). Specifically, the question to be answered was if the majority of DHSs were found upstream or downstream of the TSS. Figure 3.4 shows a peak of DHSs centred on the TSS for all cell-types and experimental conditions. Results also revealed a slight drop of DHSs slightly after +1000 bp and -1000 bp of the TSS that increased towards the +2500 and -2500 bp.

Further characterization of DHSs looked at DHS size between genomic locations to identify if significant size differences occur with DHS location. A Kruskal-Wallis rank sum test found that the location of DHSs significantly affected the size of DHSs ($p < 0.001$). A post-hoc analysis using the Dunn test identified which genomic locations were significantly different. In all cell-types and experimental conditions, upstream 1000 bp DHSs were significantly wider than other regions ($p < 0.05$). The 3'UTR (untranslated region), the downstream 200 bp, and the intergenic regions contained the next largest DHSs. The 5'UTR, exon, and intron regions contained the smallest DHSs (Figure 3.5).

Gene ontology analysis was performed to identify enriched gene functions within each cell-type and experimental condition (Figure 3.6). Gene ontology analysis was restricted to genes containing DHSs within their upstream 1000 bp. The highest enriched category for every dataset was found to be 'response to stress'; closely linked to the third highest, response to abiotic or biotic stimulus.

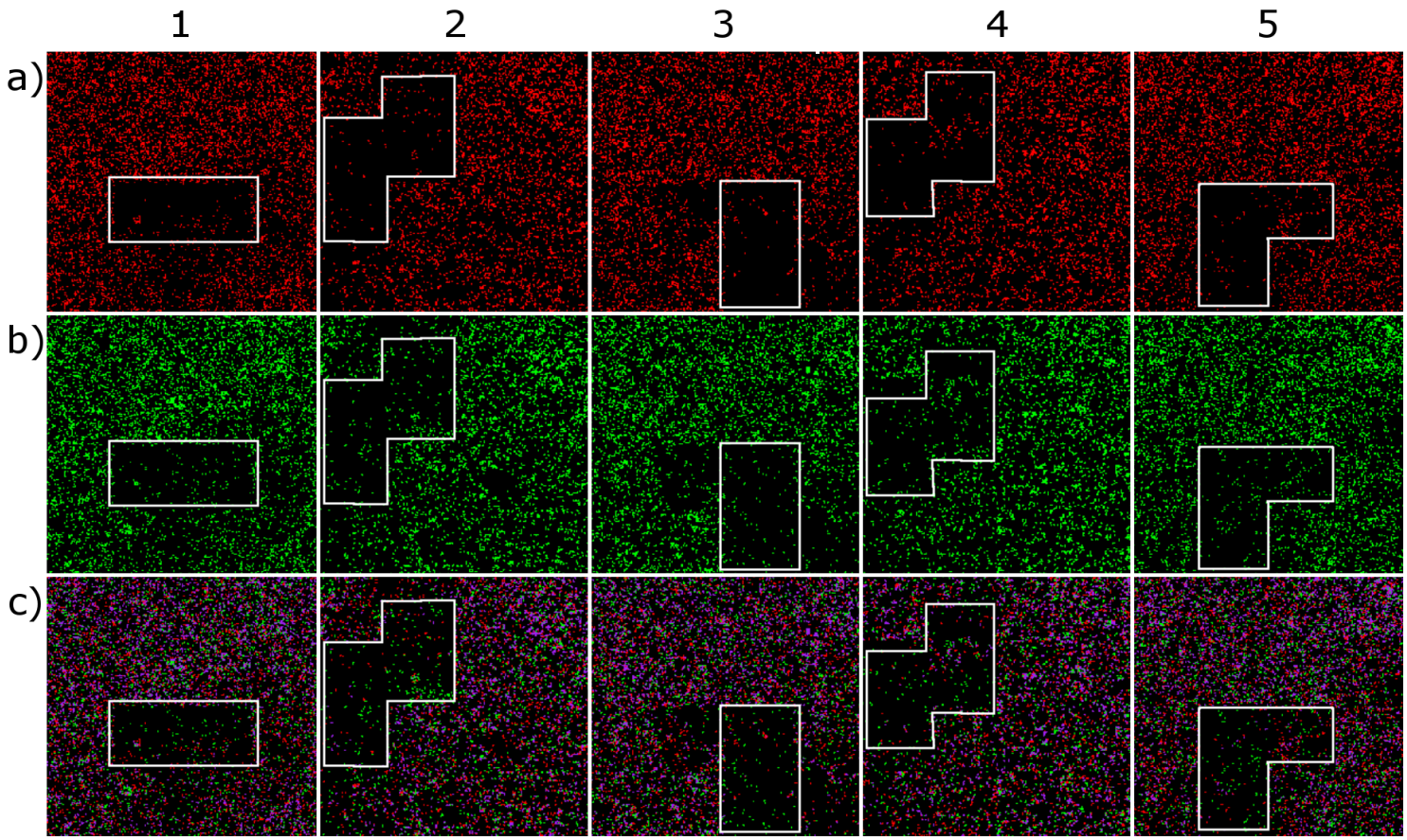


Figure 3.3: **Hilbert plots showing the distribution of DHSs across each chromosome for each experimental condition and their intersection.** Chromosome numbers for each column are indicated above the figure. Each coloured dot within the image represents a DHS. Centromeres are located within the white highlighted regions. a) Distribution of epidermal DHSs (red) b) Distribution of endodermal DHSs (green) c) Intersection of epidermal and endodermal DHSs shown in purple.

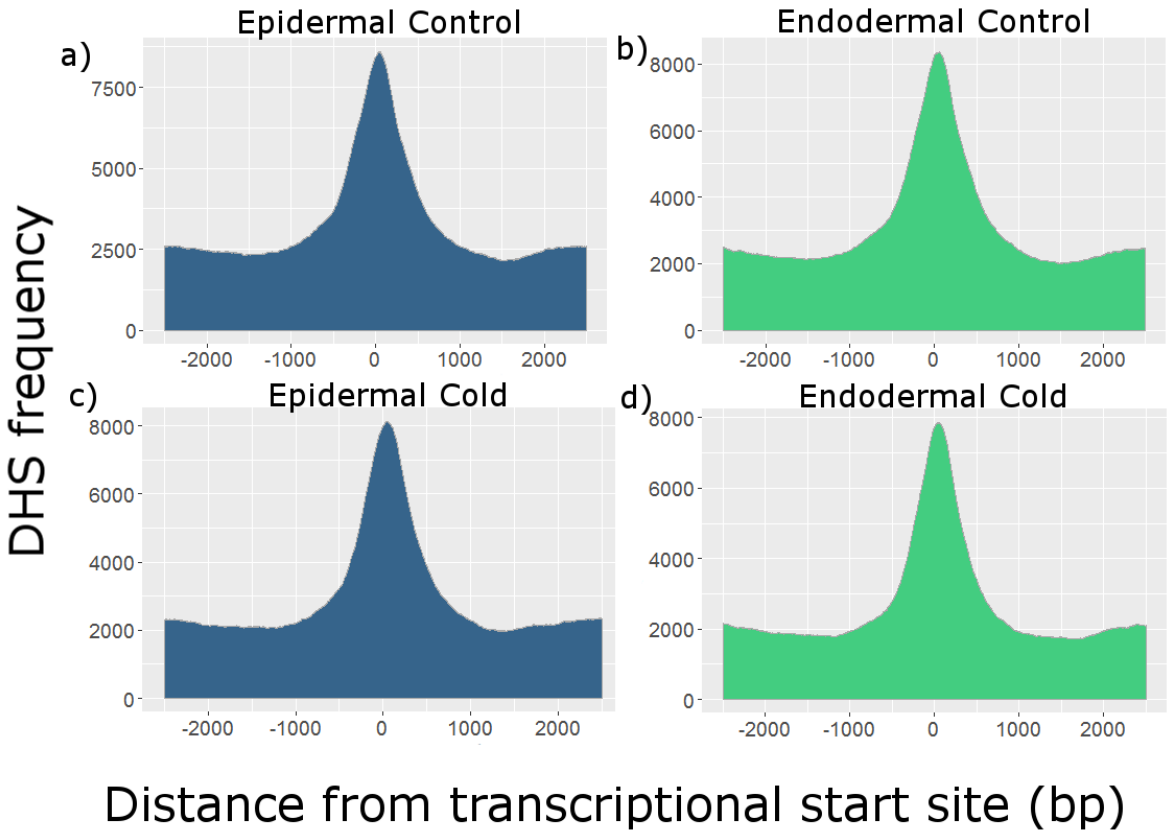


Figure 3.4: **Mean frequency of DHSs 2500 bp upstream and downstream of the transcriptional start site.** DHS abundance is centred over the TSS. a) Epidermal control b) Endodermal control c) Epidermal cold acclimated d) Endodermal cold acclimated.

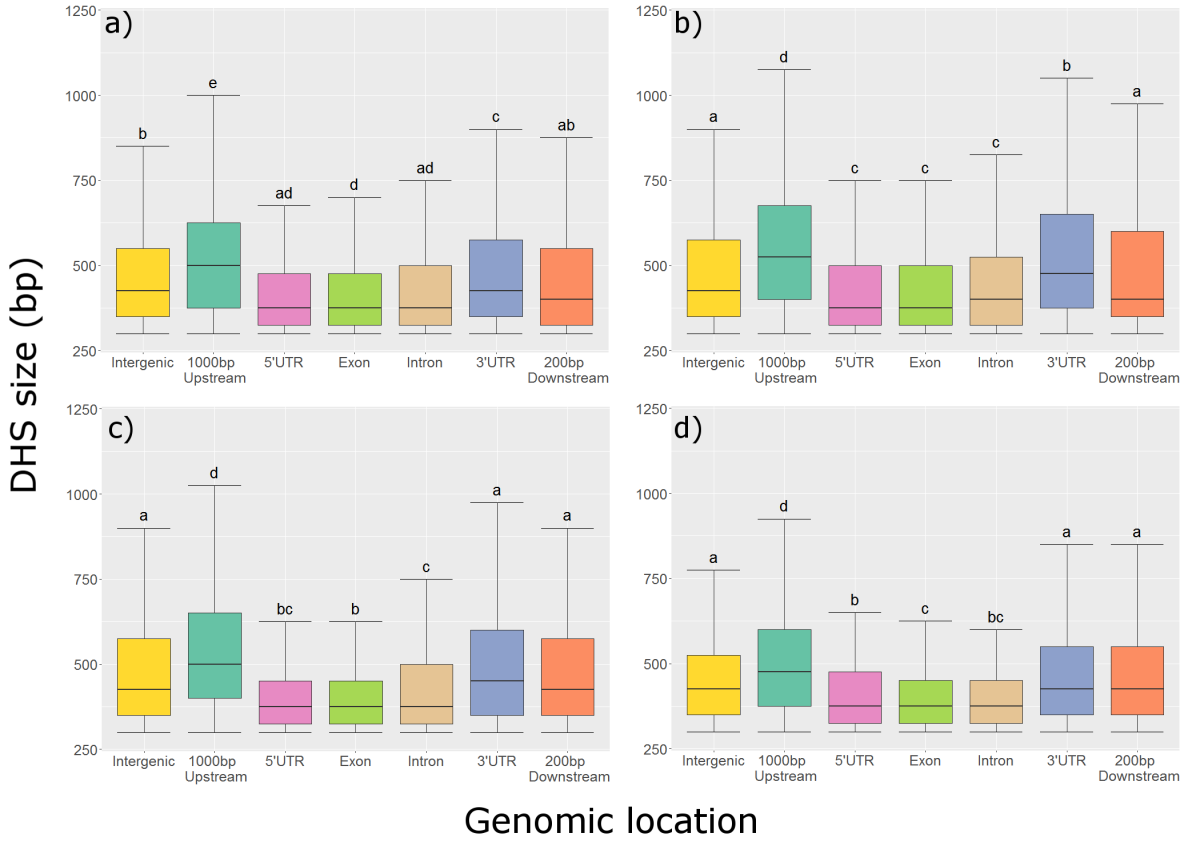


Figure 3.5: **Box plots of DHS size in base pairs with respect to genomic location.** Genomic locations in order are intergenic, 1000 bp upstream of TSS, 5'UTR, exon, intron, 3' UTR, and 200 bp downstream of translational termination site. Note that all plots start at 300 bp due to a minimum DHS size of 300 bp. Genomic locations with the same letter are not significantly different (Kruskal-Wallis & Dunn test, $p < 0.05$). a) Epidermis control b) Endodermis control c) Epidermis cold acclimated d) Endodermis cold acclimated

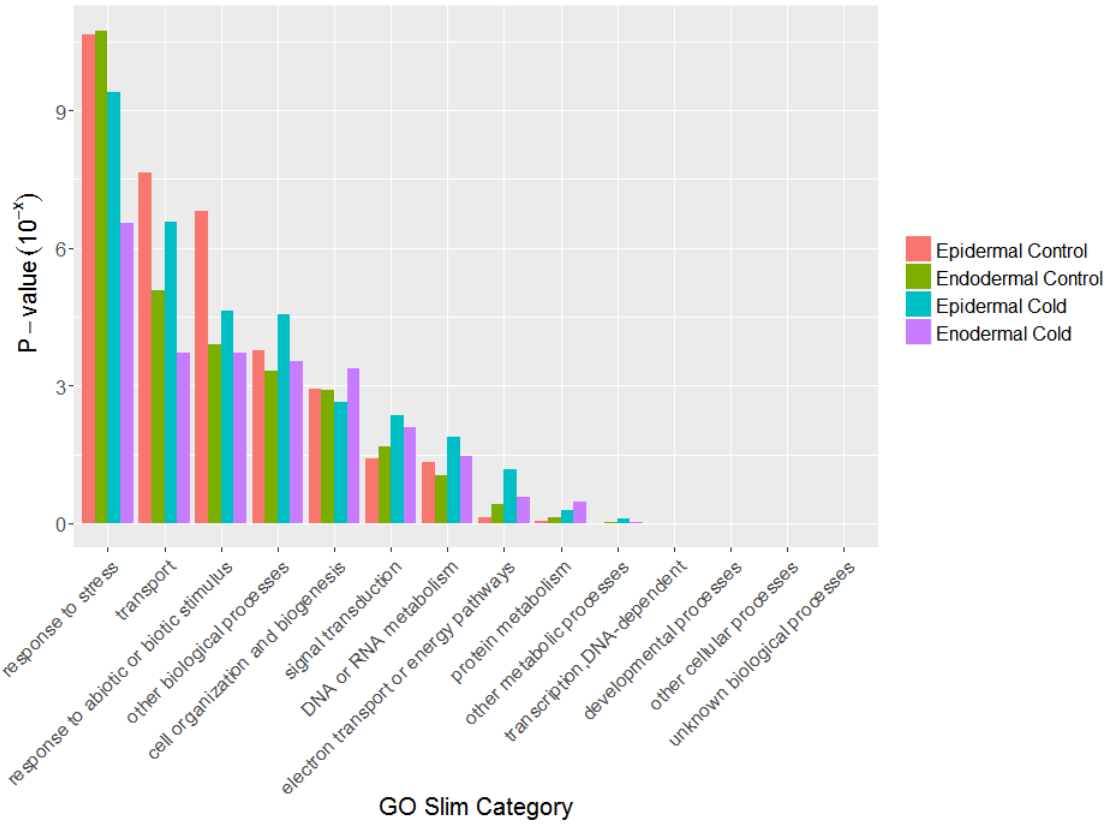


Figure 3.6: **Gene ontology slim analysis of upstream 1000 bp DHSs in each experimental condition.** Shown on the x-axis are the gene ontology slim categories and on the y-axis the associated p-value (10^{-x}) from the hypergeometric test.

3.4 DHSs correlate with distinct transcriptional patterns

Results showed DHSs were highly localized within the upstream 1000 bp and enriched around the TSS. Likewise, DHSs were found significantly larger within the upstream 1000 bp. These results point to the potential that DHSs are significantly altering transcriptional output by affecting the accessibility of a gene's promoter. To identify if DHSs are correlated with gene expression levels, the epidermal and endodermal DHS datasets were integrated with existing RNA-seq data. Raw RNA-seq data was obtained from Li et al. (2016) in which RNA from the root epidermis and endodermis were isolated through fluorescence activated cell sorting. For use in this work, RNA from Li et al. (2016) was obtained from the NCBI SRA database under BioProject PRJNA323955. Raw reads were mapped to the Arabidopsis TAIR10 genome and the FPKM for each gene was calculated for integration with DHS data.

Figure 3.7 displays the percentage of genes with a DHS in each genomic category across the highest and lowest expressed genes. For the purposes of this figure, any DHS falling within the 5'UTR and the upstream 1000 bp was not removed from the 5'UTR category. This was to identify if the 5'UTR DHSs display a similar trend to the 1000 bp upstream, since DHSs were found centred on the TSS. The epidermal and endodermal integrated data display nearly identical results. The percentage of genes with a upstream 1000 bp DHS and a 5'UTR DHS display a decreasing trend from the highest to lowest expressed genes. However, this observation is due to the upstream 1000 bp DHSs overlapping the 5'UTR region and the fact that DHSs localize centrally over the TSS. Removing 5'UTR DHSs crossing the upstream 1000 bp eliminated this effect. Thus this trend is the result of TSS DHSs. In comparison, the percentage of genes with a DHS in the intron and 3'UTR display a slight decreasing trend. DHSs within the exon and downstream 200 bp do not display any specific decreasing or increasing expression trend.

The highest expressed genes possess the highest percentage of DHSs with 80.13% of genes in the epidermal dataset and 77.64% of genes in the endodermal dataset containing a DHS in at least one genomic location. In comparison, 29.02% of the weakest expressed epidermal genes possess a DHS and 24.63% of lowest expressed endodermal genes possess a DHS. This

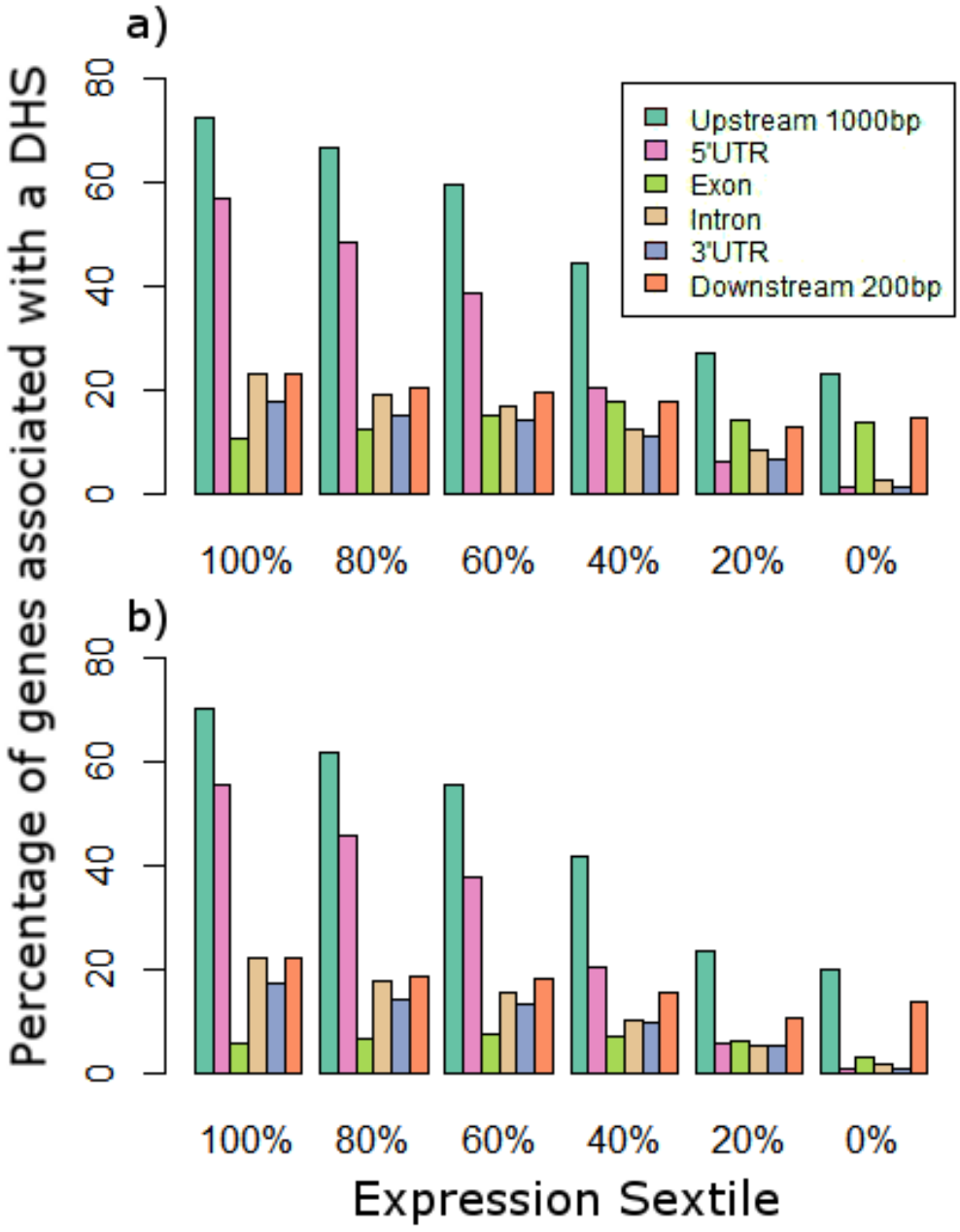


Figure 3.7: **Percentage of genes containing a DHS from the epidermal and endodermal control samples across genomic locations and expression levels.** Genes were divided into six equal sized bins from highest (100%) to lowest (0%) expression from three replicates. RNA-seq data obtained from Song *et al*, 2016. DHSs were classified into genomic locations if the DHS overlapped the region by at least 1%. a) Epidermal DHSs associated with epidermis RNA-seq data b) Endodermal DHSs associated with endodermis RNA-seq data.

analysis was repeated with cell-type specific microarray data obtained from Birnbaum et al. (2003) and obtained similar results (data not shown due to redundant results).

With the previous decreasing trend in mind, it was expected this trend would be observed in the cold acclimated datasets. However, to date there are no cell-type specific RNA-seq studies on cold acclimation. Instead, the cold DHS data was compared with total root microarray data from cold acclimated Arabidopsis (Hannah et al. (2005)). Only general trends can be taken from this analysis as the data was collected from different tissue and growing conditions. DHS data compared with cold microarray data displayed the decreasing trend observed in the control datasets (Figure 3.8). However, since the microarray data are on the whole root under different experimental conditions, the association appears weaker compared with the control datasets. Despite this, a decreasing trend is observed from highest to lowest expressed genes in the upstream 1000 bp, the 5'UTR, and in the intron genomic locations. The 3'UTR displayed a weak decreasing trend in the cold datasets. Compared with the control data only 70.73% of the highest expressed genes in the microarray data contain an epidermal cold DHS and 70.23% contain an endodermal cold DHS.

As mentioned in the introduction, genome accessibility acts like a dimmer switch in which accessibility is a continuous spectrum from completely off to completely on (Zhang et al. (2012a); Liu et al. (2017); He et al. (2012)). TF binding is significantly reduced or significantly enhanced depending on how accessible a DHS may be to TFs. The proceeding analysis tested the hypothesis that a DHS's accessibility would affect TF binding and therefore transcriptional output. It was expected that highly accessible DHSs would result in the highest expressed genes. Due to DHSs displaying an association to gene expression levels, it was also tested if a gene's expression level could be predicted based upon the DHS's accessibility.

Every DHS is identified by its respective t-test score and average likelihood ratio between the undigested and digested probability values. The likelihood ratio is an indirect measure of the digestion level or a DHS's accessibility. A DHS highly sensitive to DNase I digestion will show a high likelihood ratio between the undigested and digested sample. DHSs were sorted

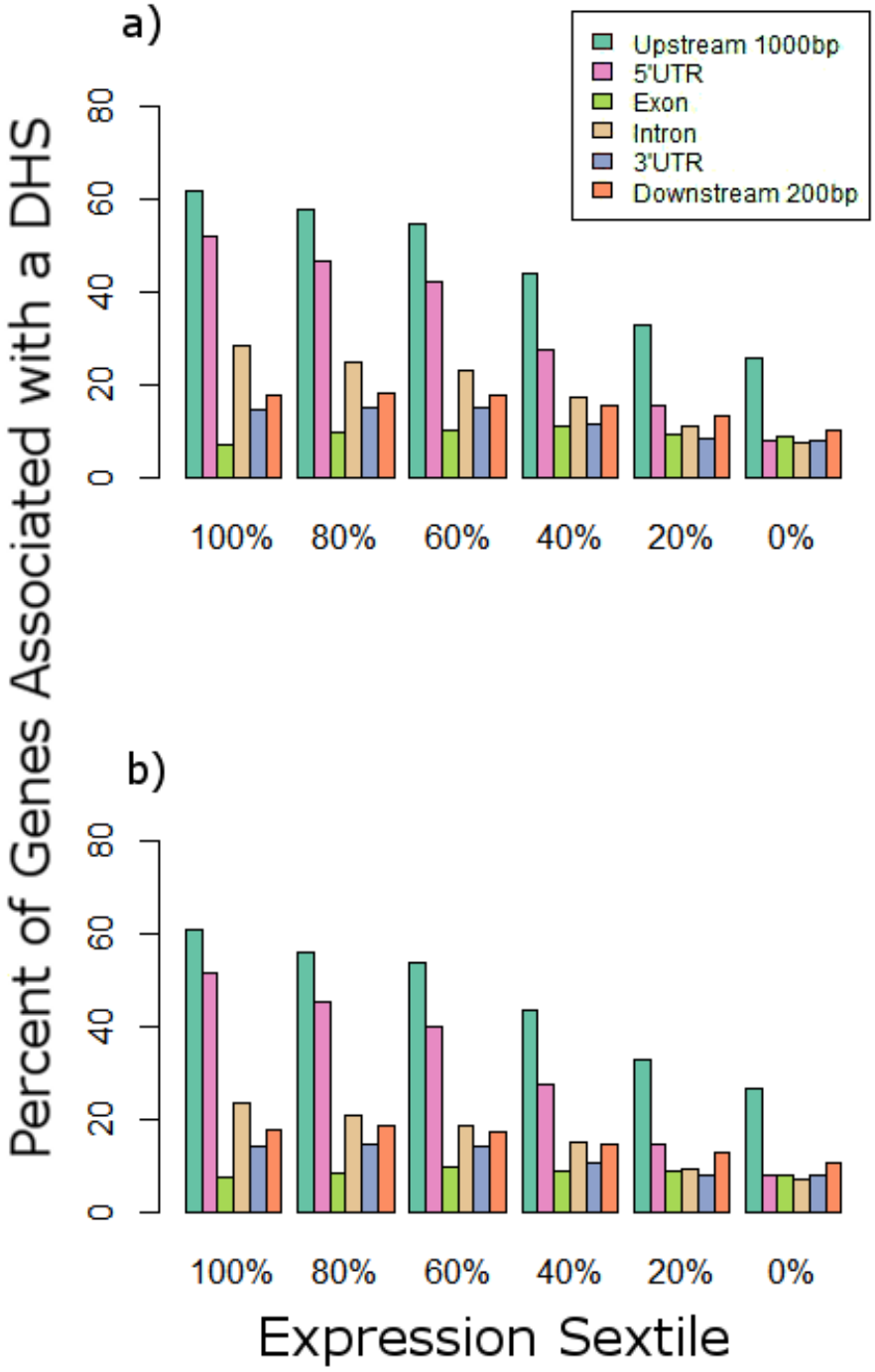


Figure 3.8: **Percentage of genes containing a DHS from the epidermal and endodermal cold samples across genomic locations and expression levels.** Genes were divided into six equal sized bins from highest (100%) to lowest (0%) expression based of the mean probe intensity of three replicates. Microarray data obtained from Hannah et al. (2005). DHSs were classified into genomic locations if the DHS overlapped the region by at least 1%. a) Epidermal cold DHSs associated with cold microarray data. b) Endodermal cold DHSs associated with cold microarray data.

separately based on the t-test score and likelihood ratio from highest to lowest and divided into six separate bins. Furthermore, DHSs were separated based on genomic location and gene expression levels. To statistically identify if an association between a gene's accessibility and FPKM existed, an ANOVA was performed for each genomic location and cell-type.

The t-score and likelihood ratio of DHSs significantly affected the FPKM of associated genes for each genomic location except 3'UTR and downstream 200 bp DHSs ($p < 0.001$). Due to the likelihood ratio more significantly impacting the FPKM of genes, it was used for the rest of the analysis. The fitted FPKM means with 95% confidence intervals for each cell-type and condition from highest to lowest likelihood ratio are displayed in Figure 3.9. Epidermal DHSs within the upstream 1000 bp ($F=57.2$ on 5 Df, $p < 0.001$), 5'UTR ($F=41.24$ on 5 Df, $p < 0.001$), exon ($F=82.77$ on 5 Df, $p < 0.001$), and intron ($F=17.56$ on 5 Df, $p < 0.001$) had a significant effect on the FPKM of their associated genes. An identical effect was observed for endodermal DHSs within the upstream 1000 bp ($F=53.36$ on 5 Df, $p < 0.001$), 5'UTR ($F=28.83$ on 5 Df, $p < 0.001$), exon ($F=44.3$ on 5 Df, $p < 0.001$), and intron ($F=8.536$ on 5 Df, $p < 0.001$). However, DHSs within the 3'UTR and the downstream 200 bp did not have a significant effect on the FPKM of their genes ($F=1.732$ on 5 Df, $p = 0.124$; $F=0.665$ on 5 Df, $p = 0.650$ - $F=0.5$ on 5 Df, $p = 0.777$; $F=1.52$ on 5 Df, $p = 0.180$). Genes with a DHS in the upstream 1000 bp, 5'UTR, exon, or intron displayed a decreasing FPKM from the highest to lowest likelihood ratio (Figure 3.9). Genes with a DHS in their 5'UTR had the highest average FPKM of all genomic locations (Figure 3.10). Genes with a DHS in their intron had the second highest average FPKM of all genomic locations (Figure 3.10). Despite DHSs in 3'UTR regions affecting transcriptional output in Figure 3.7, the accessibility of DHSs in 3'UTR regions did not increase or decrease the FPKM (Figure 3.9 b,h; Figure 3.10). In addition, while exon presence did not affect transcriptional output in Figure 3.7, the accessibility of exon DHSs significantly altered gene FPKM.

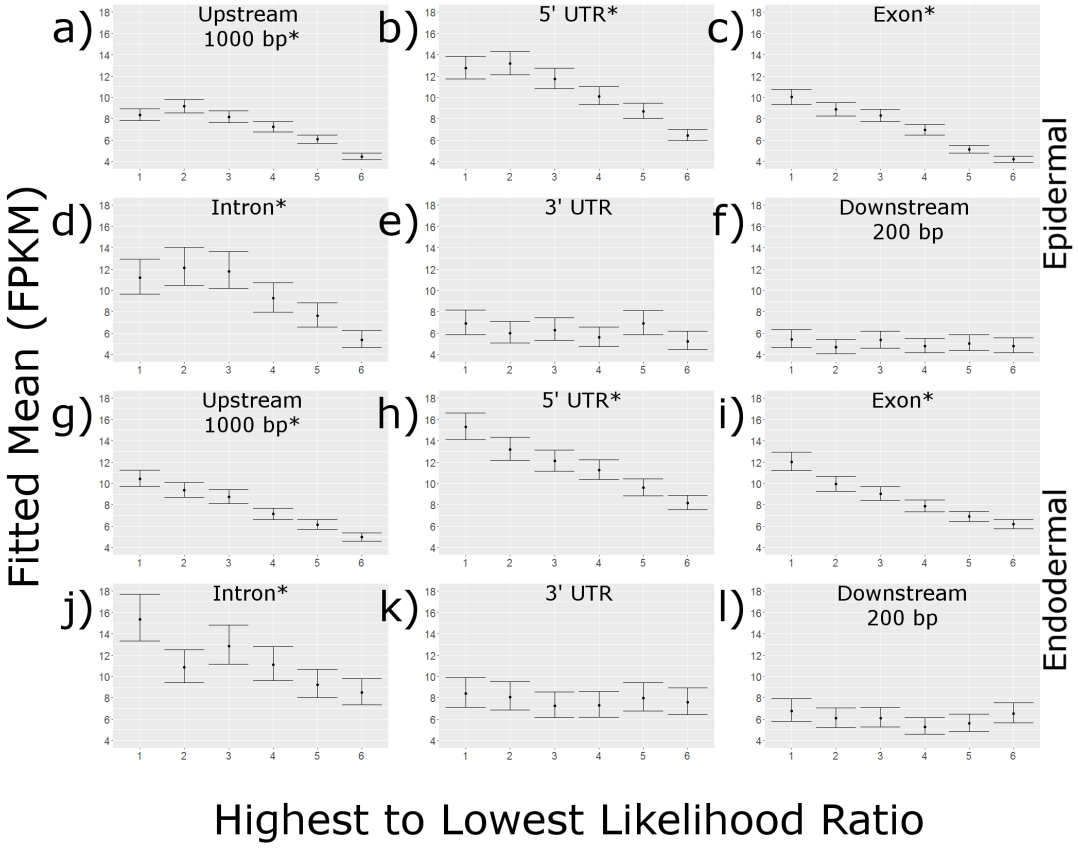


Figure 3.9: **Fitted mean of fragments per kilobase of transcript per million reads mapped (FPKM) with 95% confidence intervals across genomic locations separately.** DHSs in each genomic location were sorted based on their fold likelihood ratio from highest (1) to lowest (6). Significant effects ($p < 0.001$) are indicated with *. a-f) Epidermal DHSs with epidermal RNA-seq . g-l) Endodermal DHSs with endodermal RNA-seq. a,g) Upstream 1000 bp. b,h) 5'UTR. c,i) Exon. d,j) Intron. e,k) 3'UTR. f,l) Downstream 200 bp.

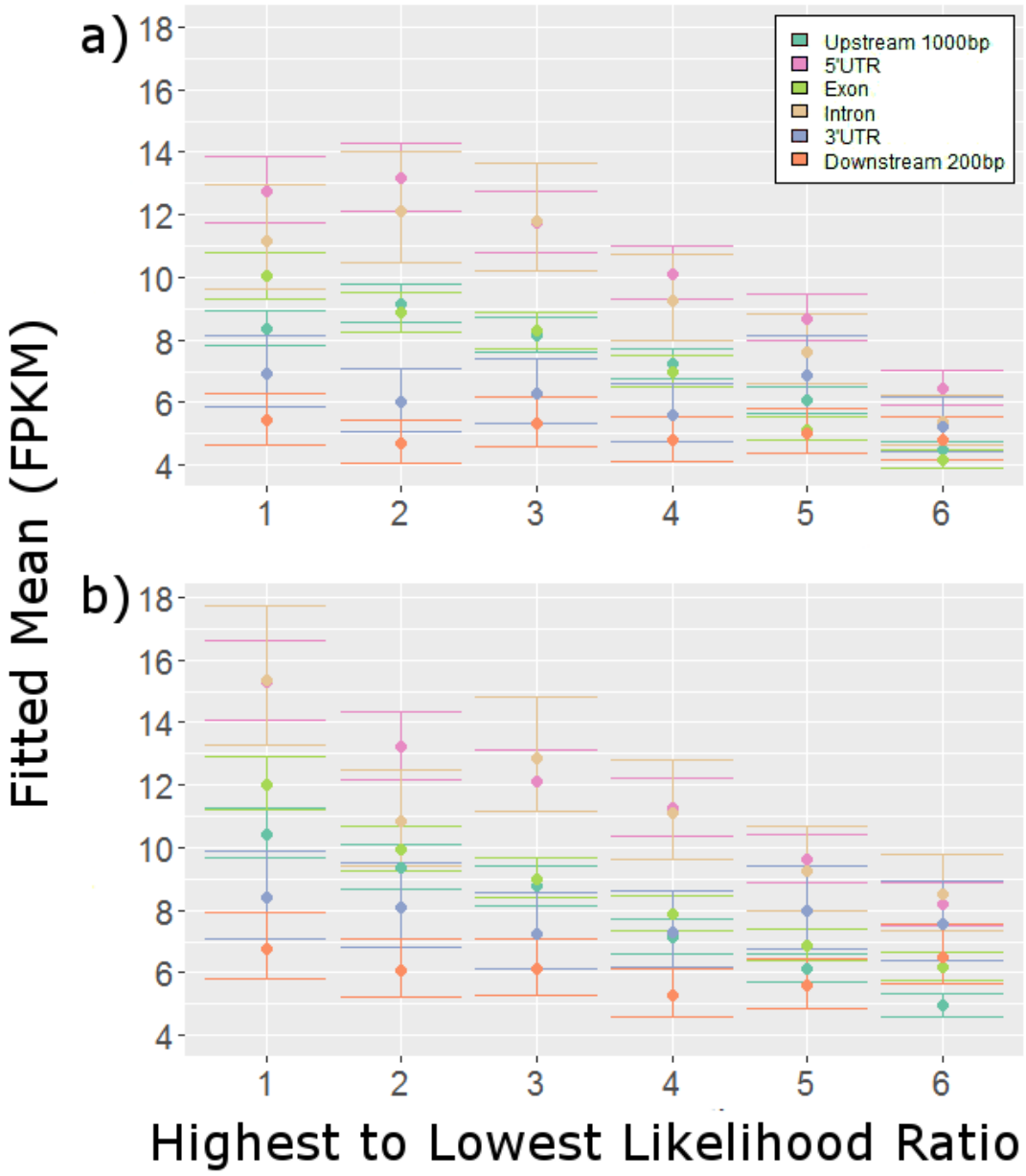


Figure 3.10: **Fitted mean of fragments per kilobase of transcript per million reads mapped (FPKM) with 95% confidence intervals across genomic locations.** DHSs in each genomic location were sorted based on their fold likelihood ratio from highest (1) to lowest (6). a) Epidermal DHSs with epidermal RNA-seq. b) Endodermal DHSs with endodermal RNA-seq

3.5 DHSs correlate with distinct epigenetic patterns

Genome wide methylation patterns from raw and processed Methyl-seq data were analyzed to identify methylation patterns around DHSs in the Arabidopsis root (Kawakatsu et al. (2016)). Here, fluorescence activated cell sorting was used to isolate GFP tagged nuclei from the root epidermis and endodermis utilizing the *WEREWOLF* and *SCARECROW* gene promoter. Bisulfite converted DNA were sequenced on an Illumina HiSeq 2000, mapped to the 'cytosine to thymine' converted Arabidopsis TAIR 10 reference genome, and methylated cytosines identified through a binomial distribution. A detailed description of all methods used, including read mapping and identification of methylated cytosines, can be found in Kawakatsu et al. (2016) and Lister et al. (2009).

3.5.1 CpG, CHG, and CHH methylation display distinct patterns

For each cell-type and condition the average methylation percentage 2500 bp upstream and downstream of each DHS centre were identified, calculated, and plotted from identified methylated cytosines (Figure 3.11). Results found CpG and CHG methylation decreased within DHSs for each cell-type. CpG methylation displayed a slight increase in methylation 500 bp before and after endodermal DHSs (Figure 3.11b). However, this increase was short as 1000 bp upstream and downstream of DHSs the CpG methylation decreased once more. CHH methylation across DHSs demonstrated a unique trend in which epidermal DHSs displayed a very sharp and distinctive increase in methylation. A similar spike was observed in endodermal DHSs, however, the increase appeared less noticeable as the epidermal methylation spike. A decrease in methylation was observed prior to the peak, as was observed in the epidermal dataset.

As shown in Figure 3.4, DHSs are highly clustered around gene TSSs. With this in mind, the previous methylation trends could be indicative of a TSS feature rather than a DHS feature. To exclude such an artifact, the methylation percentage was identified and plotted across DHSs

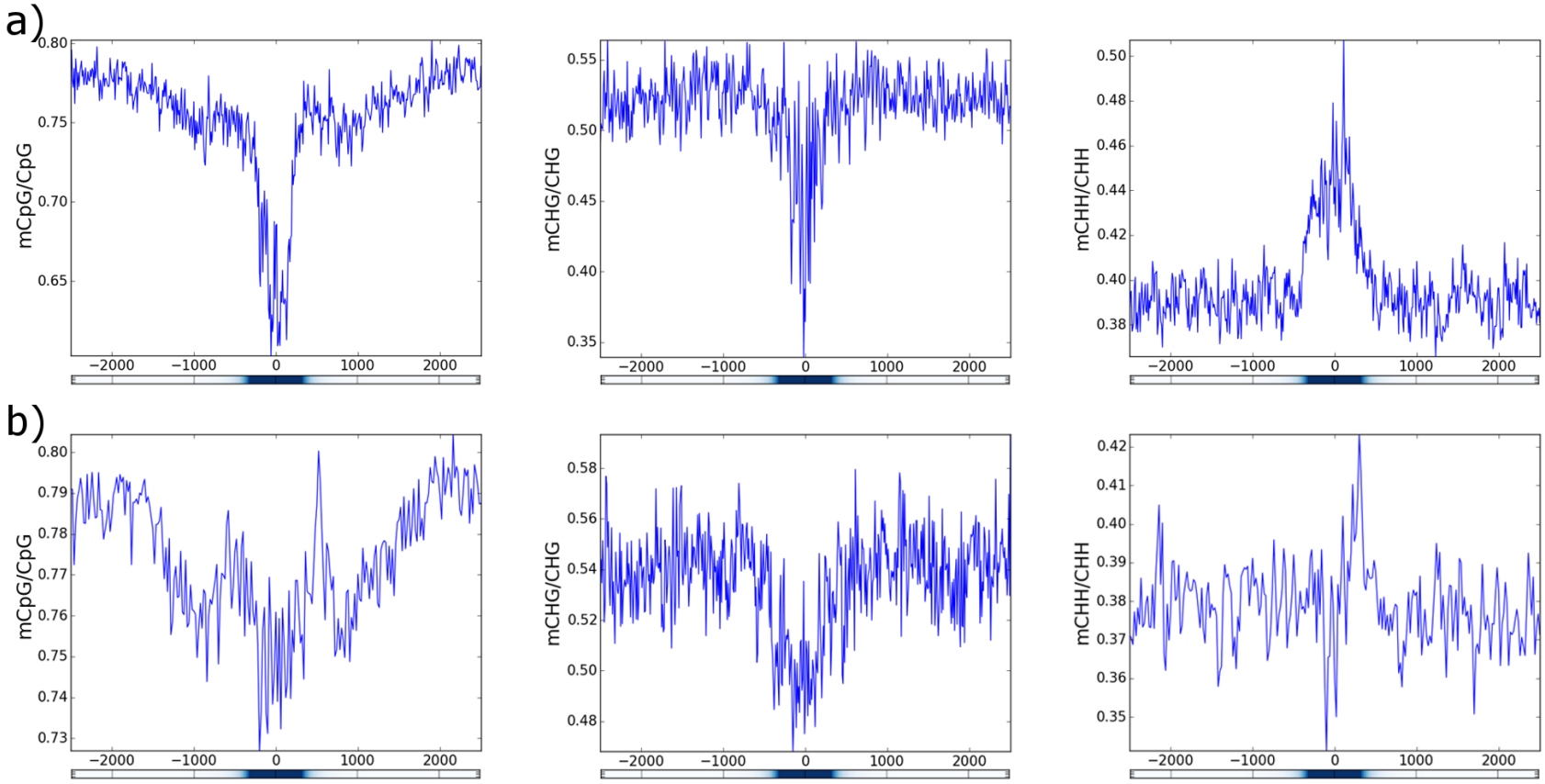


Figure 3.11: **DNA methylation patterns 2500 bp upstream and downstream of all DHSs.** Only methylated cytosine positions were used in this analysis. mCpG, mCHG, and mCHH mean methylation levels are shown as the ratio of methylated cytosine reads to all cytosine reads. Position 0 on the x-axis denotes the centre of DHSs. DHS distribution from the 0th position are shown below each subplot. a) Mean methylation levels across all DHSs for epidermal control b) Mean methylation levels across all DHSs for endodermal control.

that do not occur at the TSS (Figure 3.12). The decrease in CpG and CHG methylation within DHSs is still present and similar to results with TSS DHSs. However, the decrease observed 1000 bp upstream and downstream, in addition to the increase in methylation 500 bp upstream and downstream, were not present. CHH methylation within the epidermal DHSs retained the distinctive methylation increase. Likewise, CHH methylation within the endodermal DHSs retained the increase in methylation and the decrease in methylation observed prior to this spike.

The previous analysis utilized identified methylated cytosines but missed information from unmethylated cytosines. As a result, the previous analysis only included information on statistically methylated cytosines and misses methylation information from cytosine positions that were not significantly methylated. To that end, the analysis was repeated with the methylation percentage from all cytosines as long as the cytosine had at least four mapped reads. This increased the number of CpG methylation sites mapped around epidermal DHSs from 576,045 to 3,741,065. Furthermore, it increased mapped CHG methylation sites from 148,107 to 3,820,633 and CHH methylation sites from 376,244 to 18,769,608. CpG methylation sites around endodermal DHSs increased from 563,972 to 386,717. Lastly, CHG sites increased from 185,581 to 3,940,192 and CHH sites from 322,821 to 19,238,648. Performing an analysis with this cytosine information allows the methylation profile across all cytosines to be considered.

Figure 3.13 shows the methylation percentage around all DHSs in the epidermal and endodermal datasets. Despite distinct similarities between this analysis and the previous, many slight differences are observed. For instance, while CpG methylation is observed to sharply decrease within DHSs for both cell-types, the decrease appears more gradual than the previous analysis. Unlike the previous analysis, CHG methylation displayed a small increase around DHSs before it decreased within DHSs. This effect is observed within both the endodermal and epidermal DHSs. Furthermore, a large decrease in CHH methylation is observed within DHSs rather than a peak of CHH. In other words, while CHH methylation increased within DHSs for

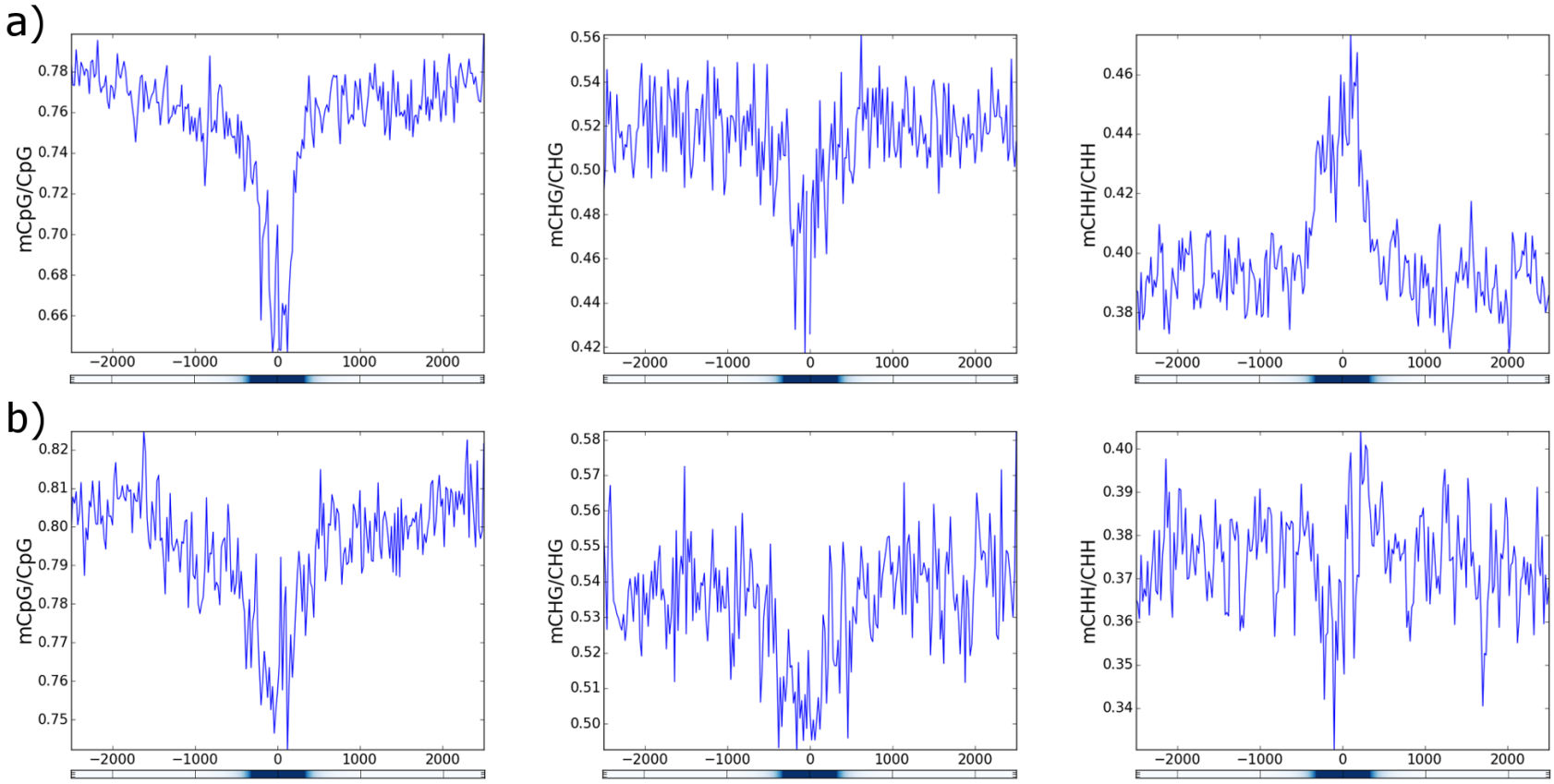


Figure 3.12: **DNA methylation patterns 2500 bp upstream and downstream of DHSs that do not cross the TSS.** Only methylated cytosine positions were used in this analysis. mCpG, mCHG, and mCHH mean methylation levels are shown as the ratio of methylated cytosine reads to all cytosine reads. Position 0 on the x-axis denotes the centre of DHSs. DHS distribution from the 0th position are shown below each subplot. a) Mean methylation levels across DHSs for epidermal control b) Mean methylation levels across DHSs for endodermal control.

methylated cytosines, CHH methylation in a broader aspect decreases. Finally, a prominent increase in CHH methylation is observed around the sharp decrease in CHH methylation within DHSs (Figure 3.13).

To remove a potential TSS artifact, the previous analysis was repeated with all cytosines by mapping methylation around a DHS dataset lacking TSS DHSs. Figure 3.14 displays the methylation level, using all cytosines, around a DHS dataset lacking TSS DHSs. Identical results were observed for this analysis compared to the previous analysis. A decrease in methylation percentage within DHSs is observed for CpG, CHG, and CHH methylation. CHG methylation increased downstream of DHSs to levels higher than upstream of the DHS. In addition, an increase is observed for CHH methylation downstream of DHSs that is not seen prior to DHSs in both the epidermal and endodermal datasets.

3.5.2 Histone modifications display modification-specific patterns

To identify histone modification patterning around DHSs, genome wide histone data were mapped 2500 bp upstream and downstream of DHSs. Processed H3K4me3 and H3K27me3 data were obtained from Deal and Henikoff (2010). Here, H3K4me3 and H3K27me3 ChIP DNA were hybridized to a custom Roche NimbleGen microarray alongside H3 ChIP DNA from the same biological sample. H3 ChIP DNA was hybridized for normalizing nucleosome occupancy during computational analysis. Histone data for both H3K4me3 and H3K27me3 were mapped and averaged across DHSs for both epidermal and endodermal DHS data (Figure 3.15). H3K27me3 +500 bp and -500 bp from DHSs increased followed by a decrease within DHSs back to background levels (Figure 3.15a,c). Likewise, H3K4me3 +500 bp and -500 bp from DHSs increased. However, rather than decrease back to background levels, a third spike in H3K4me3 was observed directly within DHSs (Figure 3.15b,d).

To remove a potential TSS artifact, the previous analysis was repeated with a dataset lacking TSS DHSs (Figure 3.16). While H3K27me3 decreased within DHSs and increased 500 bp upstream and downstream of DHSs, similar to previous results, the downstream 500 bp increase

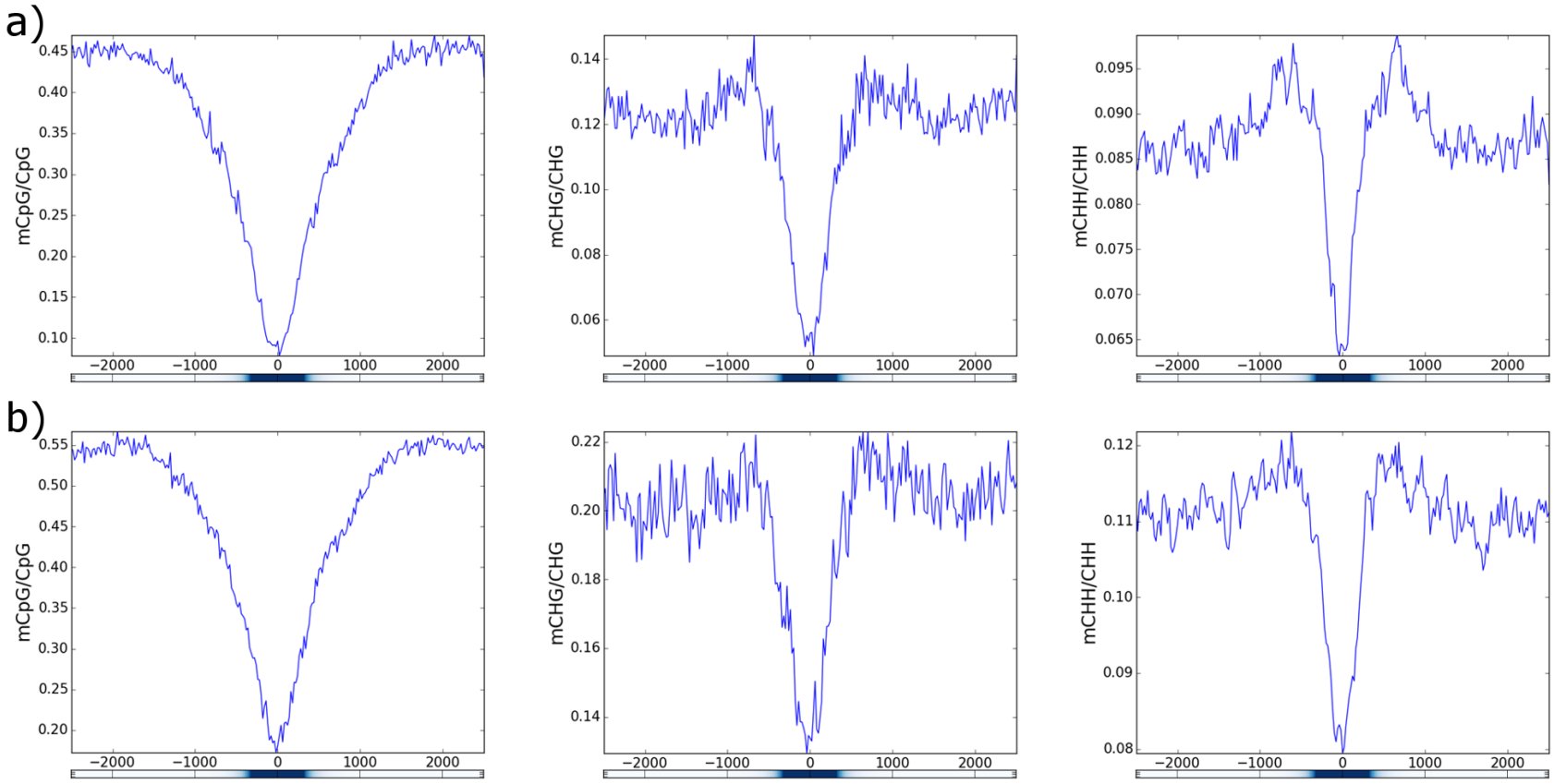


Figure 3.13: **DNA methylation patterns 2500 bp upstream and downstream of all DHSs.** All cytosine positions were used in this analysis regardless of methylation status. mCpG, mCHG, and mCHH mean methylation levels are shown as the ratio of methylated cytosine reads to all cytosine reads. Position 0 on the x-axis denotes the centre of DHSs. DHS distribution from the 0th position are shown below each subplot. a) Mean methylation levels across all DHSs for epidermal control b) Mean methylation levels across all DHSs for endodermal control.

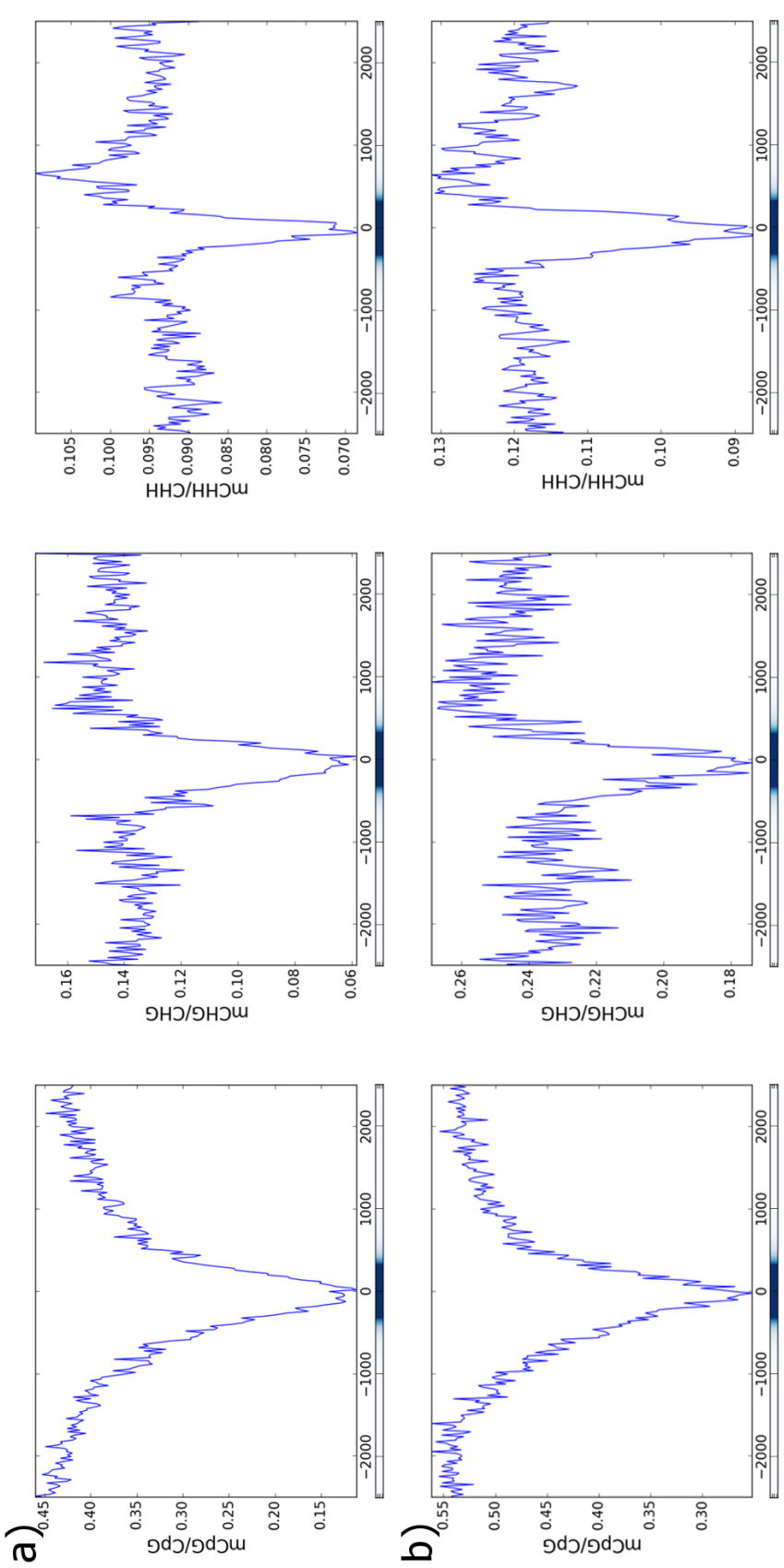


Figure 3.14: **DNA methylation patterns 2500 bp upstream and downstream of DHSs that do not cross the TSS.** All cytosine positions were used in this analysis regardless of methylation status. mCpG, mCHG, and mCHH mean methylation levels are shown as the ratio of methylated cytosine reads to all cytosine reads. Position 0 on the x-axis denotes the centre of DHSs. DHS distribution from the 0th position are shown below each subplot. a) Mean methylation levels across DHSs for epidermal control b) Mean methylation levels across DHSs for endodermal control.

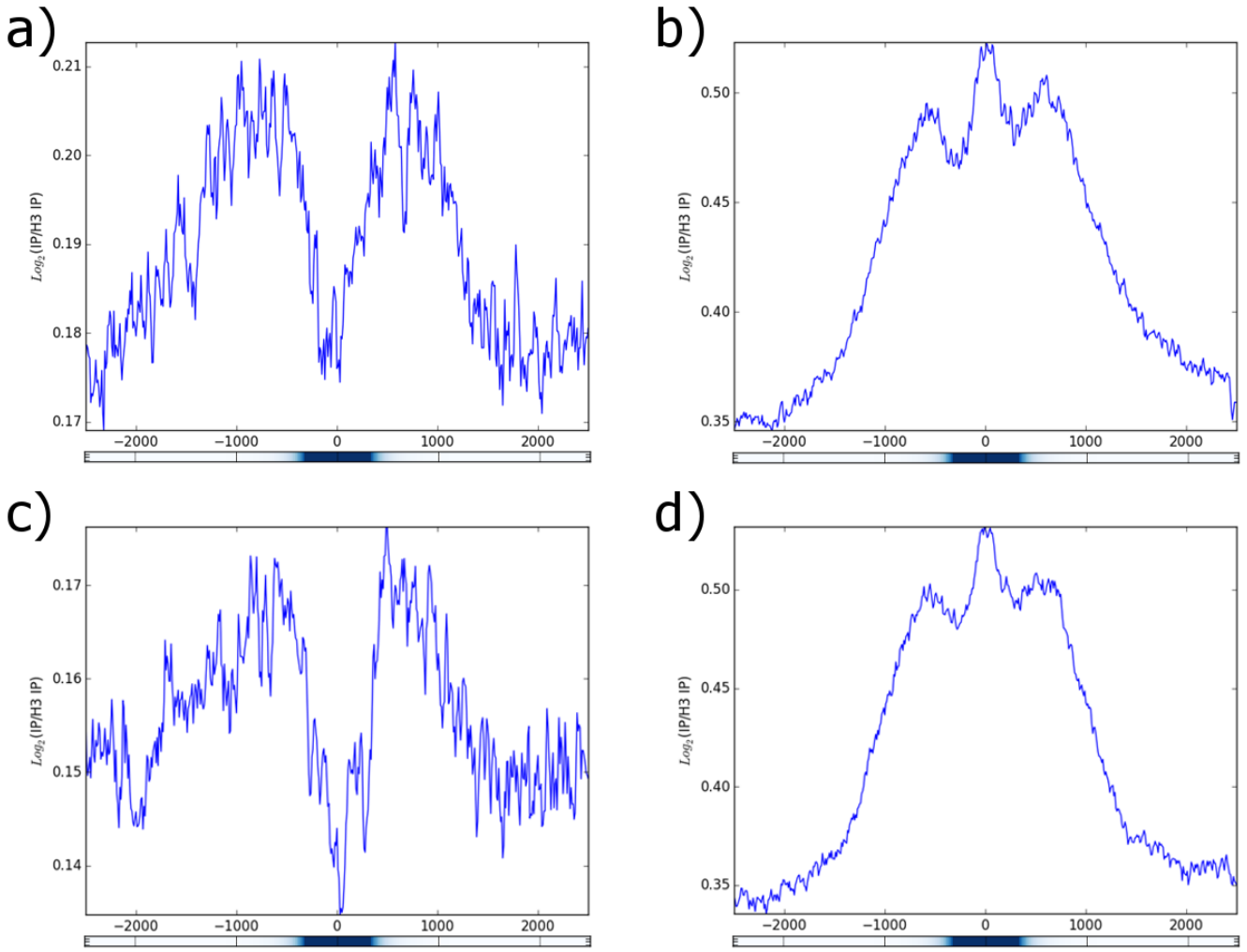


Figure 3.15: **Histone modification patterns 2500 bp upstream and downstream of all DHSs.** Standard deviate log-ratios of H3K4me3 and H3K27me3 to H3 ChIP DNA are obtained from the average of two biological replicates. DHS distribution from the 0th position are shown below each subplot. a) H3K27me3 chromatin landscape around all epidermal control DHSs. b) H3K4me3 chromatin landscape around all epidermal control DHSs. c) H3K27me3 chromatin landscape around all endodermal control DHSs. d) H3K4me3 chromatin landscape around all endodermal control DHSs.

is greater than the upstream increase (Figure 3.16a,c). In addition, results identified H3K4me3 decreased within DHSs rather than increase as in the previous analysis. (Figure 3.16b,d). However, a slight increase in H3K4me3 is observed -500 bp from the centre of DHSs.

3.6 *A priori* motif enrichment finds unique TF binding patterns

To further connect DHSs with a transcriptional response, *a priori* motifs identified through previous ChIP-seq experiments were tested for enrichment in selected gene groups from the epidermis, endodermis, and cold regulated pathways (Weirauch et al. (2014)). An incentive of integrating DHS data with motif data is to filter out motifs with the potential to be biologically active. Motifs within accessible regions are open to TF binding and therefore are biologically active, while motifs within inaccessible regions are closed to TF binding and biologically inactive. In other words, identifying motifs within DHSs serves as a form of biological validation. Furthermore, through identifying motifs within DHS, the motifs responsible for chromatin remodelling may be identified. Lastly, the motifs responsible for cell identity and stress response will be identified by mapping motifs within gene groups from the epidermis, endodermis, and cold regulated pathways. Coupled with DHS data, motif mapping will enable the selection of motif targets for biotechnology applications and future mutagenesis experiments.

A priori motifs were obtained from a study by Weirauch et al. (2014) in which known Arabidopsis transcription factors were used to generate an expansive motif library through high-throughput ChIP-seq experiments. This motif library is used to map TF binding sites and test for motif enrichment within the promoter regions of selected gene groups. The assessment of motif enrichment within selected gene groups was performed using *Cismer* on the upstream 750 bp and downstream 250 bp genomic sequences from the TSS for all genes (Austin et al. (2016)). Enriched motifs were summarised by transcription factor family and mapping locations plotted in selected gene groups.

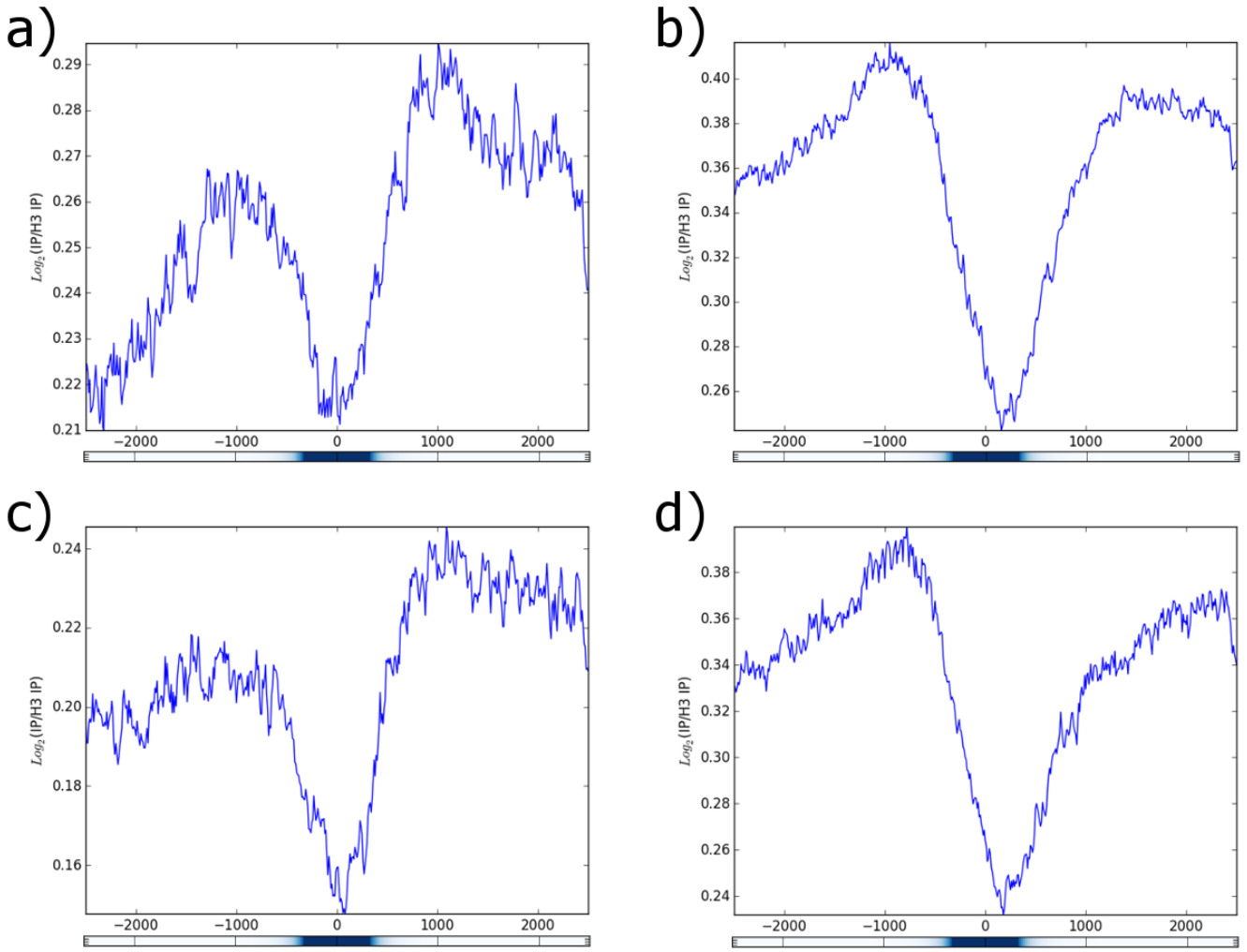


Figure 3.16: **Histone modification patterns 2500 bp upstream and downstream of DHSs that do not cross the TSS.** Standard deviate log-ratios of H3K4me3 and H3K27me3 to H3 ChIP DNA are obtained from the average of two biological replicates. DHS distribution from the 0th position are shown below each subplot. a) H3K27me3 chromatin landscape around epidermal control DHSs. b) H3K4me3 chromatin landscape around epidermal control DHSs. c) H3K27me3 chromatin landscape around endodermal control DHSs. d) H3K4me3 chromatin landscape around endodermal control DHSs.

3.6.1 Epidermal and endodermal enriched motifs

Prior to assessment of motif enrichment in the epidermis and endodermis, selected gene groups from each category were identified. To accomplish this, epidermal and endodermal upregulated genes were identified through comparison of publicly available epidermal and endodermal RNA-seq data (Li et al. (2016)). From this, 793 genes were found to be upregulated in the epidermis and 462 genes were found to be upregulated in the endodermis ($p < 0.01$). Additionally, genes with a DHS specific to the endodermis or epidermis in the upstream 1000 bp were identified. To have high confidence in these gene lists, a DHS had to have at least 50% of its length within the upstream 1000 bp in order for the gene to be added. 2112 genes had a cell-type specific epidermal DHS in their upstream 1000 bp. 1501 genes contained a cell-type specific endodermal DHS in their upstream 1000 bp. Differentially-expressed (DE) differentially-accessible (DA) genes were identified by combining the cell-type specific gene lists from the RNA-seq analysis and the DHS analysis. From this 54 DE/DA epidermal genes and 64 DE/DA endodermal genes were identified (B). *A priori* motifs from Weirauch et al. (2014) were mapped to these gene groups and tested for significant enrichment compared to a randomized background.

Figure 3.17 displays the locations of enriched motifs within the 54 epidermal DE/DA genes along with their DHSs. Four enriched motifs categories were identified in the epidermal promoters. Enriched motifs belonged to the transcription factor families of AT-hook ($Z > 3.35$), Dof ($Z = 3.45$), homeodomain ($Z > 3.58$), and general transcription binding proteins (TBP) ($Z > 3.12$). Interestingly, all of the enriched binding motifs are rich in adenine and thymine basepairs (AT rich sequences). Lastly, the TF binding sites were highly enriched in the upstream 750 bp and depleted in the downstream 250 bp.

Figure 3.18 displays the locations of enriched motifs within the 64 DE/DA endodermal genes along with identified DHSs. Two enriched motif categories were identified in the endodermal promoters: the AT-hook ($Z > 3.16$) and the homeodomain transcription factor family ($Z > 3.01$). These two transcription factor families also have enriched motifs within the epi-

Figure 3.17: **Arabidopsis epidermal DE/DA genes mapped with *a priori* motifs and DHSs from the epidermal cell layers.** Shown are the upstream 750 bp and downstream 250 bp from the TSS. Gene names and AGIs are labelled on the left for each gene. White regions are identified DHSs.

dermal gene list and similarly have rich AT binding domains. Similar to the epidermal motif distribution, a high proportion of these binding motifs are located within the upstream 750 bp.

3.6.2 Cold pathway enriched motifs

A database of upregulated cold acclimated genes was obtained from Hannah et al. (2005). 672 long term upregulated genes (upregulated >72 hours) were identified from this database and used for motif enrichment analysis. Of these, 447 genes contained an upstream 1000 bp epidermal DHS and 410 contained an upstream 1000 bp epidermal cold DHS. 343 of the 672 genes contained an epidermal DHS in both control and cold conditions, indicating the majority of genes are consistently open. 41 of the shared genes contained an additional unique upstream epidermal control DHS and another 31 shared genes contained an additional unique upstream epidermal cold DHS.

A list of 9304 epidermal cold genes contained an epidermal cold DHS overlapping the promoter by at least 50%. This list was further reduced by removing genes containing an

Figure 3.18: **Arabidopsis endodermal DE/DA genes mapped with *a priori* motifs and DHSs from the endodermal cell layers.** Shown are the upstream 750 bp and downstream 250 bp from the TSS. Gene names and AGIs are labelled on the left for each gene. White regions are identified DHSs.

epidermal control DHS. From this, 1897 genes contained an epidermal cold DHS but lacked an epidermal control DHS in the upstream 1000 bp. In contrast, 3776 genes contained and epidermal control DHS but lacked an epidermal cold DHS. In other words, 1897 genes became accessible and 3776 genes closed due to cold stress in the epidermis. Intersecting the 1897 genes with the 672 long term upregulated genes resulted in 53 DE/DA epidermal cold genes (Appendix B).

A priori motifs from Weirauch et al. (2014) were assessed for enrichment within the 53 DE/DA epidermal cold genes. Motifs for the transcription factor families of AP2 (Z>3.26), bZIP (Z>3.04), SQUAMOSA binding proteins (SBP)(Z=3.33), MADS box (Z=3.04), and several unknown (Z=3.00) were enriched in the epidermal cold DE/DA genes. The locations of DHSs and enriched motifs within this epidermal cold DE/DA genes are displayed in Figure 3.19. MADS box binding domains were identified to be heavily distributed downstream of the TSS in the 5'UTR while all other enriched motifs were distributed upstream of the TSS.

Repeating the previous analysis on endodermal DHS data found similar results. 432 of the 672 upregulated cold genes contained an endodermal upstream 1000 bp DHS and 399 contained an endodermal cold upstream 1000 bp DHS. 328 of the 672 upregulated genes contained a DHS in both control and cold conditions. 28 of the shared genes contained an additional unique upstream endodermal DHS and another 28 contained an additional unique endodermal cold DHS. Similar to the epidermal dataset, a list of cold endodermal genes containing a DHS overlapping the promoter by at least 50% were selected. Any genes containing a DHS in the control sample were removed. 2296 genes contained an endodermal cold DHS but lacked an endodermal control DHS in the upstream promoter. In addition, 5833 genes contained an endodermal control DHS but lacked an endodermal cold DHS. In other words, 2296 genes in the endodermis became accessible and 5833 genes closed under cold acclimation. Intersecting the 2296 accessible genes with the 672 long term upregulated genes resulted in 53 DE/DA endodermal cold genes. The transcription factor families of AP2 (Z>3.07), bZIP (Z>3.08), E2F (Z>3.19), storekeeper (Z=3.00), and MYB-SANT (Z=3.43), were enriched in the endodermal

Figure 3.19: **Arabidopsis epidermal cold DE/DA genes mapped with *a priori* motifs and DHSs from the epidermal cold dataset.** Shown are the upstream 750 bp and downstream 250 bp from the TSS. Gene names and AGIs are labelled on the left for each gene. White regions are identified DHSs.

cold DE/DA genes. Figure 3.20 shows the locations of DHSs and enriched motifs within the 53 DE/DA genes. AP2, E2F, and MYB-SANT were found heavily distributed downstream of the TSS in the 5'UTR. Storekeeper motifs were distributed across the 1000 bp range and bZIPs were distributed around the TSS evenly.

In order identify biologically active motifs within the upstream cold CBF pathway, eleven genes from the upstream cold CBF pathway were selected and assessed for *a priori* motif enrichment using Cismer (Austin et al. (2016)). Figure 3.21 displays the mapping locations of enriched motifs and DHS locations in all cell-types and experimental conditions across the 11 cold pathway genes. TF families with enriched motifs within this gene list are as follows: the AP2 (Z>3.17), bZIP (Z>3.02), Myb-SANT (Z>3.22), TBP (Z>3.2), bHLH (Z>3.00), CG-1 (Z=3.50), homeodomain (Z>3.03), and unknown (Z=4.13). The dominant core binding motif enriched within this dataset is called the G-box with the consensus sequence of CACGTG. The TF families binding to this motif are the bHLH' and the bZIP. The second enriched motif identified is known as the AP2 binding motif with the core sequence of CCGAC. More importantly, this motif is known as the DRE/C-repeat cis-acting element to which CBF proteins bind.

Figure 3.20: **Arabidopsis endodermal cold DE/DA genes mapped with *a priori* motifs and DHSs from the endodermal cold dataset.** Shown are the upstream 750 bp and downstream 250 bp from the TSS. Gene names and AGIs are labelled on the left for each gene. White regions are identified DHSs.

As seen from Figure 3.21 most of the upstream cold pathway genes contain a DHS in all cell-types and experimental conditions. However, *ICE1* was found lacking a DHS in all datasets except within the endodermal cold DHS dataset. As well, *RAP2.1* lacked a DHS under control conditions but upon cold acclimation a DHS in both the epidermis and endodermis opened. Unique observations can be made by observing this pathway and the motifs identified in them. For instance, CBF1, CBF2, CBF3 are under the control of the MYB15 protein and importantly contain a Myb-SANT binding motif downstream of the TSS. *RAP2.1*, *RAP2.6*, *LOS2*, and *ZAT10* promoters all contain multiple AP2 binding motifs. Interestingly, CBF1, CBF2, and CBF3 are AP2 transcription factors and the previously mentioned genes are under the control of CBF proteins. One additional gene, *HOS1*, contained three AP2 binding motifs. The TF family bHLH contained binding motifs dispersed across all these genes. Notably, ICE1 is included within the bHLH family. It is also interesting to note a highly densed portion of binding motifs within a DHS of *HOS1*, possibly indicating a cis-regulatory module. Lastly, for *CBF1* an upstream DHS is found to open in the endodermal and epidermal cold conditions.

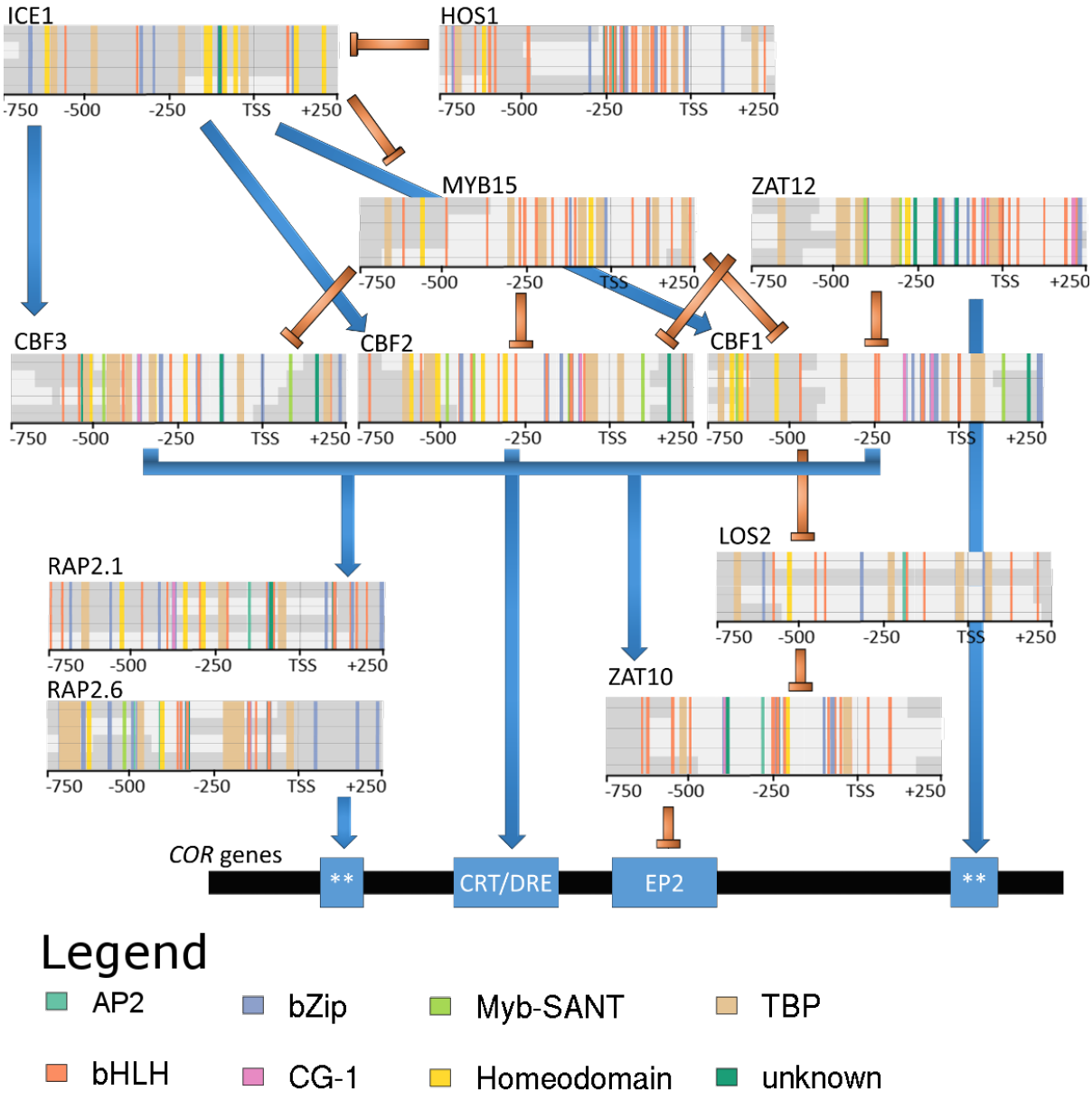


Figure 3.21: **Arabidopsis cold acclimation pathway mapped with a *priori* motifs and DHSs from the epidermal and endodermal cell layers under control and cold conditions.** Shown are the upstream 750 bp and downstream 250 bp from the TSS of the upstream genes in the cold acclimation pathway. In order from top to bottom from each gene are the DHSs from the epidermal control, epidermal cold, endodermal control, and the endodermal cold samples. White regions are identified DHSs.

Chapter 4

Discussion

The development of cell-types and tissues; how they acquire cell identity, and how they respond to environmental conditions is a question researchers have been studying for decades. Within the Arabidopsis root, cell-type expression profiles have been generated for most cell-types and tissues (Birnbaum et al. (2003); Li et al. (2016)). Despite this, how cells acquire cell-type specific transcriptional responses and the downstream effects of transcriptional differences has been largely understudied. The entire biological story from DNA to transcription, to proteomics, to metabolomics remains to be pieced together. This work focused on identifying how epigenetics, specifically chromatin accessibility, affects the transcription component of this story. This study aimed to investigate the chromatin dynamics of cell-type identity and stress response through DHS identification and integration with transcriptomic data and epigenetic data within the Arabidopsis root epidermis and endodermis cell layers.

This work developed a custom designed program, called *DDTS* paired with a novel next generation sequencing (NGS) protocol, to identify DHSs from cell-type specific populations of the Arabidopsis root and to overcome challenges of traditional DNase-seq protocols. It has been used to successfully identify thousands of DHSs from the Arabidopsis root epidermis and endodermis under control and cold conditions. Through computational assessment, DHSs displayed characteristics shared among cell-types and experimental conditions. DHSs displayed

shared genomic location distributions, sizes, frequency distribution around the TSS, distinct overlaps, and gene ontologies. Integration with transcriptomic data identified unique transcriptional patterns revealing higher expressed genes to be significantly more associated with DHSs than low expressed genes. Furthermore, gene expression levels were found statistically associated with the accessibility of their DHSs. Integration with epigenetic data identified a distinct drop of DNA methylation and histone modifications within DHSs. Lastly, *a priori* motifs were tested for enrichment within DHSs of cell-type specific and stress responsive genes. From this many TF families and respective motifs were identified from cell-type and stress responsive gene promoters. Overall, this work presents a comprehensive analysis and identification of DHSs in the Arabidopsis root epidermis and endodermis through the use of a novel DNase-seq protocol and associated computational analysis tool. A discussion of how transcription, transcription factors, TF motifs, DHSs, methylation, and histone modifications, interact to respond to external conditions and develop cell identity are discussed within the following sections.

4.1 DNase-DTS: Advancement in DNase-seq protocols and analysis

To overcome problems associated with previous DNase-seq protocols and analysis procedures, the INTACT protocol in conjunction with Nextera and *DDTS* was developed. Previous work in the Austin lab developed the wet lab protocols necessary to perform DNase-seq on cell-type isolations from the Arabidopsis root. The remaining step for this work was to develop analysis pipelines and procedures to process the generated data. The DNase-DTS protocol and analysis program proved effective at overcoming issues with previous DNase-seq protocols and in identifying DHSs at a cell-type level.

4.1.1 Challenges with existing DNase-seq studies

For DNase-seq to be performed in difficult tissues on a cell-type level, certain issues with existing DNase-seq protocols first needed to be solved. The first complicated and time consuming DNase-seq step solved was with the use of agarose gel plugs. Agarose gel plugs were required in DNase-seq protocols to prevent mechanical shearing of DNA, albeit at the cost of time and low DNA output (Boyle et al. (2008a); Crawford et al. (2006)). The low DNA output prevented DNase-seq from being used within tissues where DNA isolation was difficult. Cumbie et al. (2015) simplified DNase-seq and enabled isolation of nuclei from uncooperative Arabidopsis tissues and saved two days of wet lab work by removing agarose plugs. Despite this, the DNase-seq protocol was still time consuming and required agarose gel analysis. The agarose gel analysis is required to assess DNA digestion over a DNase I dilution series in order to select the sample optimally digested by DNase I to sequence in order to enable efficient and accurate DHS identification.

Though Cumbie et al. (2015) enabled analysis of difficult tissues, DNase-seq on single cell-types was challenging and time consuming to perform due to the low amounts of nuclei obtained in addition to high cellular debris isolated. To enable single cell-type analysis the requirement of millions of nuclei for DNase-seq protocols needed solving. Single cell-type analysis was feasible, although time consuming and tedious, if many nuclei isolations and DNase I digestions could be performed. Therefore, to obtain the required nuclei and DNA, DNase-seq requires the removal of the agarose gel analysis completely as it reduces the total DNA output (Boyle et al. (2008a); Crawford et al. (2006); Song and Crawford (2010)). Each DNase-seq protocol is limited by the amount of DNA obtained due to DNA isolation from an agarose gel DNA smear. Additionally, an agarose gel assessment can be inaccurate and is currently prone to the experimental bias of selecting the appropriate digestion level. One paper, by Jin et al. (2015), overcame the agarose gel assessment and performed DNase-seq analysis on a single cell, albeit in a more complicated and time consuming fashion than existing DNase-seq protocols. Lastly, despite small improvements, existing DNase-seq protocols still require

time consuming blunt end polishing, fragment enrichment, size fractionation, and extensive library preparation that limit DNase-seq's use on certain tissues and cell-types.

4.1.2 DNase-DTS improves upon existing DNase-seq challenges

To overcome these issues, the Austin lab developed the wet-lab portion of DNase-DTS which uses many of the same procedures in Cumbie et al. (2015). However, it skips agarose gel assessment and improves upon the nuclei isolation stage through the use of the INTACT protocol (Deal and Henikoff (2011)). Using INTACT, DNase-DTS isolates high quality cell-type specific nuclei easily and quickly without the use of special equipment. However, even with INTACT, isolating the required millions of nuclei in a single cell-type for typical DNase-seq protocols is challenging. Hence, the Austin lab shifted from the typical DNase-seq protocols and analysis procedures to develop the complete DNase-DTS protocol and analysis pipeline requiring a fraction of the nuclei needed by other methods.

Previous DNase-seq protocols involved enriching DNase I digested regions and identifying DHSs as peaks of sequencing above a specified threshold (Boyle et al. (2008a); Crawford et al. (2006); Song and Crawford (2010)). In addition, previous protocols required a minute amount of DNase I for minimal digestion of DHSs to enable accurate identification. Instead, DNase-DTS identifies regions of the genome lacking sequencing through comparing an undigested to a digested sample. To accomplish this, DNase-DTS requires more extensive DNase I digestion over a longer period of time in order for DHSs to be fully digested. Thus, to increase power and confidence in DHS identification, DNase-DTS require high depth sequencing. DNase-DTS next replaces assessing DNase I digestion on an agarose gel with downstream computational analysis and optional Bioanalyzer assessment. By skipping gel assessment, a greater amount of DNA is retained for sequencing library preparation.

Furthermore, DNase-DTS skips blunt end polishing, fragment enrichment, size fractionation, and extensive sequencing library preparation as these steps are not required for this protocol. Instead, DNase-DTS uses enzymatic library preparation (i.e. Illumina Nextera) to prepare

DNA for sequencing. With this method, undigested regions are prepared for sequencing while digested DHSs are not prepared for sequencing as they are heavily digested and lack sufficient DNA for enzymatic library preparation. Thus, it is important to adequately digest DHSs as they will be library prepared and sequenced if they are not sufficiently digested. DHSs are accordingly identified as a lack of sequencing in a digested sample compared to an undigested sample over a particular genome range. In summary, by skipping blunt end polishing, fragment enrichment, size fractionation, and laborious library preparation, DNase-DTS saves a significant amount of wet-lab and analysis time. In fact, everything up to data analysis requires only a single day of wet lab work.

This procedure not only allows identification of DHSs from single cell-types but improves upon previous protocols. By comparing a digested sample and an undigested sample, sequenced on the same sequencing cartridge, DNase-DTS removes many sequencing biases that affect final results. Through comparing the digested to the undigested sample, DNase-DTS also produces meaningful statistical values. However, no analysis tool exists to analyze the new data generated from the wet-lab portion of DNase-DTS. As a result, the computational analysis side of DNase-DTS (i.e. *DDTS*), was developed to compare and identify regions lacking sequencing.

DDTS is a custom designed program, written in Python, to handle the unique form of data generated from DNase-DTS. *DDTS* identifies DHSs through comparing an undigested sample to a digested sample and performing a statistical test to identify if the region significantly lacks sequencing in the digested sample. Through comparing two samples sequenced together, *DDTS* significantly improves upon previous analysis protocols by removing sequencing biases without complex algorithms. Furthermore, previous analysis programs computationally generate a background dataset for statistical analysis while *DDTS* creates a biological background in the form of an undigested sample (Boyle et al. (2008b)). *DDTS* is an easy to use Linux program written for use on a wide range of organisms, on multiple forms of data, and across various experiments. Theoretically, *DDTS* can be utilized for any analysis that utilizes differential read

abundance between sequencing runs.

Due to the creation of the new protocol and analysis procedure called DNase-DTS, new quality checks on sequencing quality and DNase I digestion needed development. Previous DNase-seq protocols would run samples on an agarose gel and select the optimal sample to sequence (Song and Crawford (2010)). However, this step involves agarose gel plugs, is quite time consuming, and is very inaccurate (Song and Crawford (2010); Cumbie et al. (2015)). Instead, DNase-DTS sequences all samples and computationally assesses DNase I digestion making it quite simple, quick, and accurate. A technician no longer has to visually assess library preparation or DNase I digestion, but instead receives more concise visual assessments and numerical results for those assessments. DNase-DTS also utilizes an optional step of running samples through a Bioanalyzer 2100 (Agilent) to assess digestion prior to sequencing. However, while optional it does not mean one should skip a Bioanalyzer assessment. DNase I digestion is highly prone to experimental errors and any slight changes in experimental conditions can result in faulty digested samples. To save costs on sequencing faulty digested samples, it would be advantageous to assess DNA libraries prior to sequencing.

Prior to running *DDTS* to identify final DHS datasets, each individual replicate is analysed individually to ensure proper DNase I digestion and that replicates are highly reproducible. To accomplish replicate assessment, each replicate is run through *DDTS* individually to produce information on its digestion profile. The first replicate assessment identifies if the number of DHSs increase as the amount of DNase I units increase. Similarly, a check to identify if DHS width increases with increasing DNase I units is performed. Analysis of this is accomplished with a table of values or in bar charts as seen in Figure 3.2. As the amount of DNase I increases, an increasing amount of DNA is digested resulting in wider and more identified DHSs. Increasing the amount of DNase allows heavier digestion of DHSs enabling a greater power in their identification. If these observations are missing, there could have been issues with DNase I digestion, sample preparation, or sequencing. For instance, if the number and size of DHSs at 0.1 U and 0.5 U of DNase I are similar, proper DNase I digestion or sample

preparation did not take place. Lastly, relative consistency in DHS width and numbers should be observed between individual replicates. Likewise, consistency and reproducibility between replicates is evaluated through a correlation assessment before running *DDTS* with replicates combined.

The last step in replicate assessment is to observe individual digestion profile figures in a bar plot and pie figure manner as shown in Appendix A. Here a proper digestion profile is observed if the percent and number of upstream 1000 bp DHSs increase as DNase I units increase. In addition, a drop in the percent of DHSs in other genomic regions should be observed. An increase in upstream 1000 bp DHSs is expected as DHSs, also called nucleosome free regions (NFRs), are localized directly upstream of gene TSSs (Albert et al. (2007); Schones et al. (2008)). A further description of why these trends are observed is discussed later. The size of DHSs and the number of DHSs across DNase I units can also be assessed in the bar plot portion of these figures. Using all the previous information together, one can properly assess the DNase I digestion and sample quality of individual replicates over the dilution series. Compared to assessing DNase I digestion on an agarose gel, the Bioanalyzer assessments and computational assessments of DNase-DTS results in extensive information on sample quality and DNase I digestion. Thus, DNase-DTS significantly improves upon existing DNase-seq studies.

Repeating this entire analysis over a dilution series enables an analysis of replicate quality and ensures proper DNase I digestion of all samples in addition to the selection of the optimally digested sample. If the sample digested at 0.5 U of DNase I is always selected for further analysis, comparing it to a dilution series will allow an extensive assessment of its quality and digestion that could not be obtained individually. Indeed, a dilution series allows researchers to more accurately assess and identify if any issues arise. For instance, DNase I digestion is itself a difficult procedure to maintain consistency between experiments and assess with a single sample. DNase I activity can vary considerably between isolates and even reduce in efficiency when stored. DNase-seq methods require tiny amounts of DNase I to be added to samples and any slight errors in pipetting can produce widely differing results. Indeed, even

an extra 30 seconds of DNase I treatment changes the degree of DNase I digestion. With only one digested sample, this extra degree of DNase I digestion may not be identified immediately. Additionally, human errors and experimental conditions may change, even slightly, and affect DNase I digestion and sample quality. Degraded samples, faulty DNase, over digested samples due to incorrect DNase I inactivation, etc, may all be accurately identified from the sequenced dilution series data.

The final step of *DDTS*, post replicate analysis, runs all replicates together to statistically identify DHSs shared across all replicates. Previously, DNase-seq studies on Arabidopsis failed to statistically identify DHSs shared across three replicates (Zhang et al. (2012b); Pajoro et al. (2014); Cumbie et al. (2015)). In fact, most DNase-seq studies analyzed at most two replicates. Even when three replicates were used, sequenced reads from all replicates were combined and DHSs identified from pooled data (Sullivan et al. (2014)). Previous DNase-seq studies identified DHSs by generating a continuous probability landscape using F-Seq and identifying the regions that exceeded a standard deviation above a computationally generated background mean (Boyle et al. (2008b)). Comparing a digested sample to an undigested sample to control for sequencing biases are never performed in traditional DNase-seq studies. *DDTS* improves upon this by performing a statistical test between the mean values of three replicates from a digested and an undigested sample, very akin to a typical microarray or RNA-seq study. By statistically comparing an undigested and a digested sample, sequencing biases are reduced and a tighter statistical rigour is applied to DHS identification.

As done with individual replicates, assessment of the digestion profile is repeated with final DHS datasets. The number of DHSs, size of DHSs, and percentage of DHSs across genomic locations are required to be assessed here in addition to the assessment on individual replicates. Any abnormality missed previously may arise here instead. More importantly, issues across three replicates or simply issues arising with all replicates combined may be identified here. It is at this point the final optimally digested sample is selected for further analysis. The main reason for the dilution series, in addition to quality checks, is to create options for choosing the

final optimally digested DHS dataset.

Upon initial analysis of new samples, a sample with optimal DNase I digestion is chosen. The selected sample should not be over digested, with closed chromatin digested, but should not be under digested, with many DHSs undigested. An optimal digested sample may be identified through several observations that indicate the sample is at its maximum DHS digestion. This selection is similar to picking the peak point on a parabola. For example, through observations it was found the percent of upstream 1000 bp DHSs increased up to a certain DNase I unit, followed by a decrease. Specifically, upstream 1000 bp DHSs were digested until they were saturated leading to digestion of closed regions, thereby, lowering the percent of upstream 1000 bp DHSs. This is expected as DHSs, also called nucleosome free regions (NFRs), are localized directly upstream of the TSS (Albert et al. (2007); Schones et al. (2008)). The final decision on the optimal digested sample is made through the analysis of all data available, including individual replicate assessment and DHS data from replicates combined. For this work, at 100,000 nuclei, 0.5 U of DNase I was found to optimally digest samples. It was the tipping point of the distribution of DHSs across genomic locations as 0.7 U decreased the percent of upstream 1000 bp DHSs and increased digestion in unexpected DHS regions such as the exon and the downstream 200 bp.

Once an optimal digested sample is selected at given *DDTS* settings, *DDTS* requires the selected sample to be re-analyzed with settings adjusted to reduce the false discovery rate (FDR). A 5% FDR must be maintained through adjustment of the t-score cutoff and the likelihood ratio cutoff. Previous DNase-seq studies often did not identify their FDR. However, when studies did identify a FDR, they compared their results to data obtained from a computationally generated dataset. *DDTS* also provides this option (Zhang et al. (2012a)).

In summary, both the NGS protocol and computational features of DNase-DTS significantly improve upon existing DNase-seq protocols by reducing the time and complexity of existing DNase-seq methods, allowing small samples sizes, and improving upon quality control mechanisms. Additionally, DNase-DTS introduces a new analysis pipeline that may be

utilized across various organisms and experiments. Finally, *DDTS* may be used with data other than DNase-seq data such as ChIP-seq data.

4.2 Epidermal and endodermal DHSs reveal unique and shared characteristics

One goal of this thesis was to identify the characteristics and the defining similarities and differences between DHSs across datasets. To accomplish that, several thousand DHSs were identified by *DDTS* and their locations, sizes, distributions, and numbers of DHSs were compared across datasets. In sum, this work identified that promoter DHSs are the largest and most numerous DHSs.

4.2.1 Chromatin's control on cell and tissue identity

DDTS was effective at accurately identifying thousands of DHSs in the Arabidopsis root epidermis and endodermis cell layers. Indeed, 15,000 to 18,000 DHSs were identified across both cell-types under control and cold conditions. Similarly, several previous DNase-seq experiments identified thousands of DHSs across various tissues within Arabidopsis. However, results seem to differ widely (Zhang et al. (2012b); Pajoro et al. (2014); Cumbie et al. (2015); Sullivan et al. (2014)). Zhang et al. (2012b) identified 38,290 and 41,193 DHSs in leaf and flower tissues, while Pajoro et al. (2014) identified 5,680 and 8,789 high confidence DHSs within flower tissues. Sullivan et al. (2014) identified 26,712 to 34,288 DHSs across various tissues and Cumbie et al. (2015) identified 57,000 and 79,000 DHSs within the leaf and root tissues. Lastly, Liu et al. (2017) identified 10,380 DHSs in seedlings under control conditions. Altogether, results tend to differ widely even within the same tissue. One explanation for differing results from this work is that single cell-type DHSs were from four week old plants rather than bulk tissue DHSs on one week old seedlings. Additionally, variation in results likely arose from the different methods, protocols, or analysis programs used to identify DHSs. Slightly

changing settings within any analysis program on the same raw data can often produce dramatically different results. Moreover, results can vary depending on sample quality, the total read number, and how digested DNA was prior to sequencing. These differing results reveal the necessity to have normalization and rigorous statistical tests to reproduce findings.

Previous results comparing DHSs from callus and seedling tissues found less than half (49.1%) of DHSs from callus tissue were shared with seedling tissue DHSs (Zhang et al. (2012a)). Considering these are widely different tissue types, it is not surprising they have distinct chromatin accessibility landscapes. In contrast, this work identified the majority, 63.6% epidermal and 71.4% endodermal, of DHSs were shared between cell-types. As these cell-types arise from the same root tissue, their high overlap is expected. Despite the high overlap, these results revealed distinctive differences in the chromatin landscape between cell-types of the same tissue with 30%-40% unique DHSs. The unique DHSs from each cell-type may potentially regulate their cell identity. Supporting this conclusion, Song et al. (2011) found a strong connection between chromatin accessibility and cell-type identity with 30% - 40% of DHSs shared between functionally related human cell-types. Thus, the observed chromatin landscape differences make biological sense as the root is a highly specialized structure and each cell-type serves specific functions.

A potential for distinct differences in accessibility between shared epidermis and endodermis DHSs exist. Accessibility is a continuous spectrum from completely open to completely closed (Zhang et al. (2012a); Liu et al. (2017); He et al. (2012)). It is possible a DHS may be accessible in both cell-types but may be more or less accessible in one or the other. The shared DHSs between the epidermis and endodermis cell layers may have specific differences in accessibility significantly affecting TF binding. A highly epidermal expressed gene may be more accessible in the epidermis than within the endodermis leading to higher TF binding and gene expression. For instance, Zhang et al. (2012a) found 19,628 or roughly 30% of shared callus and seedling DHSs had distinct changes in accessibility, indicating shared DHSs vary in accessibility between cell-types. Thus, cell identity may be the result of changed accessibility

in shared DHSs rather than simply the presence of DHSs. A change in accessibility between shared DHSs was not analysed within this work but is a possible future direction. Furthermore, a future experiment could identify if differentially-expressed genes are the result of changed accessibility between shared DHSs.

4.2.2 Gene promoters are highly accessible

Several clear DHS characteristics are displayed among all cell-types of this work and across diverse experiments. The majority of DHSs were present within the upstream 1000 bp with the second most abundant located in the intergenic regions. Compared to previous experiments, the distributions are similar, but the percentage of DHSs across all genomic locations differ. For example, Zhang et al. (2012b) found 45% of DHSs within the upstream 1000 bp of Arabidopsis, Sullivan et al. (2014) found 37% of DHSs within the upstream 400 bp, and lastly Liu et al. (2017) identified 47% of DHSs within gene promoters. Compared to 65% of DHSs within the upstream 1000 bp in this work, widely different results are found across diverse datasets. The percentage of intergenic DHSs is considerably varied across datasets with 15% of DHSs located within intergenic regions of Zhang et al. (2012b), 37% in Sullivan et al. (2014), 21.5% in Liu et al. (2017), and 15% in this work. However, these differences can easily be explained by each experiments method of classifying DHSs into genomic regions. For instance, this work defined the intergenic region separate from the upstream 1000 bp, while previous studies often grouped the upstream 1000 bp and intergenic regions. Furthermore, DHSs may be classified into two categories or be classified into a unique category even if the DHS overlaps two genomic categories. Of course, the observed variation may also be attributed to the differences in experimental procedures for DNase I digestion and analysis. The high percent, 65%, of DHSs within the upstream 1000 bp likely attests to the largely cis based control of Arabidopsis promoters (<1000 bp), that NFRs are highly enriched within upstream elements, and the accuracy of *DDTS* (Zou et al. (2011); Hernandez-Garcia and Finer (2014); Albert et al. (2007); Schones et al. (2008)). Indeed, the majority of Arabidopsis transcriptional regulation

is due to cis-regulatory elements within gene promoters (Zou et al. (2011); Hernandez-Garcia and Finer (2014)). Finally, the more clearly defined genomic regions, like exons and UTRs, are quite comparable across datasets and overall DHS genomic distributions are consistently shared across studies. For instance, all Arabidopsis datasets agree that the majority of DHSs are within gene promoters (Boyle et al. (2008b); Zhang et al. (2012a,b)).

The highly centred distribution of DHSs around the TSS in all DHS data of this work is not unexpected. Previous experiments in Arabidopsis and rice display this distribution within their datasets as well (Zhang et al. (2012b); Sullivan et al. (2014); Zhang et al. (2012a); Cumbie et al. (2015)). Previous distributions also support the high frequency of DHSs located in the upstream 1000 bp region. It is expected this distribution is the result of actively transcribing genes requiring the space around their TSS to be open for transcriptional machinery binding (Guertin and Lis (2013)). The transcriptional complex in combination with a whole array of TFs requires significant space for DNA binding, particularly in gene promoters (Guertin and Lis (2013)). Therefore, for gene transcription, transcriptional machinery requires DNA in the promoter to be free of obstructing histones. As a result, chromatin has a direct effect on the overall transcriptional output of cells (Guertin and Lis (2013); John et al. (2011); Bell et al. (2011)). This hypothesis was tested with results shown in Figure 3.7.

The DHS sizes displayed in this work are quite comparable to previous findings. In fact, previous literature reports highly variable DHS sizes with a reported average between 300 bp - 600 bp (Natarajan et al. (2012); Boyle et al. (2008a)). Similarly, this work reports averages between 400 bp-500 bp depending on the genomic location (Figure 3.5). Furthermore, the size of DHSs across genomic locations further supports the role of DHSs directly influencing transcription. Indeed, DHSs within the upstream 1000 bp genomic regions were significantly larger than other DHSs. This observation is not new as similar results were obtained by Boyle et al. (2008a) and Natarajan et al. (2012). The highly positioned large DHSs surrounding gene TSSs are necessary for active transcription as TF binding requires significant space (Pugh and Tjian (1991); Kim et al. (2005)).

4.3 Transcriptional control through chromatin accessibility

As chromatin accessibility has been previously found to inhibit TF binding and gene expression, this work set out to identify if the DHSs identified within this work could be associated with gene expression (Zhang et al. (2012a); Boyle et al. (2008a); Song et al. (2011); Felsenfeld (1992)). Thus, the thousands of DHSs identified in this work were integrated with RNA-seq data obtained from Liu et al. (2017). Results identified a clear association between the presence and the degree of DHS accessibility with gene expression.

4.3.1 DHS presence dictates gene expression level

Results displayed a high frequency of large DHSs within the promoters and TSSs of genes, indicating a role of chromatin structure in transcription. Gene expression is initiated from TF binding within the promoter and at gene TSSs. Thus, the trends and patterns of DHSs within gene promoters could not be coincidence. To explain these results, it was hypothesized that chromatin accessibility significantly affects gene transcription by limiting TF and transcriptional machinery binding. To test this, RNA-seq data was integrated with the DHS data obtained in this work (Li et al. (2016)). Results showed the percent of genes containing a DHS in the upstream 1000 bp or 5'UTR regions decreased from the highest to lowest expressed genes (Figure 3.7). The decreasing trend indicates that highly expressed genes are most associated with TSS DHSs. In other words, high gene expression requires an accessible TSS. Similarly, previous DHS datasets across various species identified the same decreasing trend, supporting these results (Zhang et al. (2012a); Boyle et al. (2008a); Song et al. (2011)). Transcription requires accessible TSS chromatin as nucleosome structure inhibits transcriptional initiation by preventing the binding of the preinitiation complex and of general transcription factors (Felsenfeld (1992); Workman and Kingston (1998); Yuan et al. (2005)).

Altogether, these results indicate that high gene expression requires accessible DHSs in the promoter and TSSs of genes. All results support the hypothesis of actively transcribed genes

requiring accessible promoter regions for transcriptional machinery and transcription factor binding. Hence, the presence of TF binding, in addition to the chromatin state of TF binding sites, should be considered when studying gene transcription (Kouzarides (2007)).

Interestingly, not all highly expressed genes in this analysis contained a DHS around their genomic location. Biologically, all highly expressed genes would be expected to contain a DHS in their promoter region. While this may be due to experimental error, there are other possible explanations worth exploring. Firstly, despite the RNA-seq data being obtained through the use of the same *WEREWOLF* (*WER*) gene promoter, FACS was used for cell-type isolation (Bonner et al. (1972); Li et al. (2016)). Thus, different cell populations could have been obtained resulting in slightly differing results. Secondly, plants were grown in completely different conditions for different lengths of time compared with this work. Notably, this project grew plants for four weeks in liquid media compared to one week on a Petri dish, potentially altering transcriptional output. Lastly, previous datasets also did not obtain 100% of their highest expressed genes containing a DHS (Zhang et al. (2012a); Boyle et al. (2008a); Song et al. (2011)). One explanation supported by Boyle et al. (2008a), puts unannotated TSSs as the blame for this observation. Specifically, many genes may have a mislabelled TSS interfering with the observed results. Additionally, Boyle et al. (2008a) gives a list of other reasons explaining this observation including: gene regions not easily sequenced, other genomic annotation errors, biological variation, and false positive rates.

Biologically, this observation of highly expressed genes not containing a DHS may simply be explained through biological exceptions. While nucleosomes significantly inhibit TF binding, many exceptions have been discovered where TFs bind even when their cis-element was enclosed within a nucleosome (Albert et al. (2007); Yuan et al. (2005); Adams and Workman (1995)). For instance, the *PHO5* gene promoter interacted with its binding element Pho4 prior to any chromatin alterations (Adkins et al. (2004)). Indeed, many pioneering factors, those that initiate changes in chromatin accessibility, bind to closed chromatin (Zaret and Carroll (2011)). An additional explanation may be due to rapid spontaneous unwrapping of nucleosomes allow-

ing temporary access to binding sites that may not be detected through DNase-DTS (Bucceri et al. (2006)). If specific binding sites rapidly change states from open to closed it may be quite difficult to detect them. Lastly, many mechanisms control gene expression and it would be too simple for chromatin accessibility to be the ultimate deciding factor. Regardless, chromatin accessibility appears to play a dominant role in gene expression.

Intriguingly, a large percentage of the lowest expressed or repressed genes contained a DHS within their upstream 1000 bp. Results showing the same observation have previously been reported (Zhang et al. (2012a); Boyle et al. (2008a)). These results beg the question as to why transcriptionally inactive genes require accessible chromatin. One possible explanation labels these genes as transcriptionally poised (Gross and Garrard (1988); Muse et al. (2007)). In other words, these genes are not currently actively transcribed but are prepared for immediate active transcription. The quick activation of the necessary genes would be advantageous for an organism requiring an immediate response to external conditions. Thus, maintaining genes in an accessible chromatin state where TF binding is only required will save significant time. Indeed, altering chromatin accessibility is a slow time consuming process and is the rate limiting step in transcriptional response (Raser and O'Shea (2004); Barbaric et al. (2001)). Additionally, transcription is controlled through many processes including TF binding, histone modifications, and DNA methylation (Jiang (2015); Vavouri and Elgar (2005); Prokhortchouk and Defossez (2008); Kass et al. (1997); Kouzarides (2007)). DHSs may require activation or changes in the mentioned processes in order for active transcription to take place.

4.3.2 DHS accessibility dictates gene expression level

The previous section emphasized that the simple presence of DHSs highly influenced gene expression levels. However, further integration of DHS data with RNA-seq data also revealed a strong correlation between the overall degree of DHS accessibility and gene expression (Figure 3.9). Results identified that high gene expression was associated with a high degree of DHS accessibility as measured by the likelihood ratio. The likelihood ratios of upstream 1000

bp, 5'UTR, exonic, and intronic DHSs were found to be directly proportional to the expression level of associated genes (Figure 3.9).

Thus, the more accessible a gene's DHS, the higher its potential expression. In support of this, previous experiments have also found DHS accessibility associated with gene expression levels (Boyle et al. (2008a); Zhang et al. (2012a). Overall, this indicates that highly accessible DHSs enable easier TF binding, thereby increasing gene expression. The more accessible a DHS is, the more likely a TF will bind. With this information it is possible to predict a gene's expression level, at least somewhat, using the genomic location of DHSs around it and DHS sensitivity. Predicting a gene's expression level based on open chromatin information has previously been attempted (Natarajan et al. (2012)). Thus, a new potential for predicting gene expression level and patterns may be utilized through DHS mapping. In summary, the degree of DHS accessibility significantly affects the potential of TF binding and thereby affects gene transcription.

A secondary conclusion from these analyses was that 5'UTR DHSs and intronic DHSs result in the highest gene expression levels of any DHSs. Biologically, a gene with a DHS in their 5'UTR allows transcriptional machinery to bind to the TSS and initiate transcription. Furthermore, genes with a DHS in the 5'UTR are more likely currently actively transcribed. The intronic DHSs resulting in higher levels of transcription is very interesting. Previous studies identified that when introns were removed from a gene, its expression level was greatly reduced (Gruss et al. (1979)). In fact, introns were found to significantly increase a gene's expression level in many different species (Duncker et al. (1997); Lu and Cullen (2003); Rose et al. (2016); Rose (2002)). Taking that into consideration, DHSs within intronic sequences may allow enhancer elements to become accessible to their binding elements, thereby enhancing gene expression (Bianchi et al. (2009); Levy-Wilson et al. (1992)). Despite an intronic DHS link to gene expression, analysis into a general mechanism has not been studied.

In summary, the presence and accessibility of DHSs significantly affects the overall ability of TF binding and hence, significantly affects the overall transcriptional output (Zhang et al.

(2014)). As a result, the creation of these epigenetic maps, i.e. locations of DHSs and chromatin structure maps, for cell-types and abiotic conditions will be essential for the complete understanding of gene expression. Research has moved passed the simplistic view of gene expression resulting from simple TF binding an activation. Rather it is now at the point where gene expression is an acknowledged complicated process involving the interaction of TF binding and an array of epigenetic mechanisms.

4.4 Methylation and histone modifications: their role with DHSs

A key question regarding epigenetics is how the individual factors of chromatin accessibility, histone modifications, and DNA methylation interact. What causes chromatin accessibility across tissue types and what are the defining factors or characteristics of open chromatin in regards to other epigenetic factors? Are there key defining factors dictating accessible chromatin from inaccessible chromatin? This work identified many defining epigenetic factors of open chromatin and DHSs.

4.4.1 DNA methylation

CpG, CHG, and CHH methylation displayed unique patterns within and around DHSs (H is an adenine, cytosine, or thymine). DHSs were found hypomethylated in CpG and CHG contexts in both cell-type datasets. Similar hypomethylation within DHSs was identified within Arabidopsis and other organisms, supporting results (Zhang et al. (2012a); Sullivan et al. (2014); Thurman et al. (2012)). As DHSs are associated with active transcription, a lack of methylation within DHSs makes biological sense. Specifically, CpG methylation is associated with transcriptional silencing while hypomethylation is linked to transcriptional activation (Siegfried et al. (1999); Vining et al. (2012); Thurman et al. (2012)). Methylation blocks transcription through multiple processes with the end result of interfering with TF binding (Jones et al.

(1998); Klose and Bird (2006)). Therefore, in order to activate transcription, hypomethylation within DHSs is required for transcription factor binding. In other words, active transcription appears to require both open chromatin and hypomethylated DNA for efficient TF binding. Supporting this, the link between DNA methylation, DHSs, and TF binding was previously reported (Wiench et al. (2011)).

The most interesting finding from the methylation analysis was the spike in CHH methylation within DHSs using the limited set of CHH methylated cytosines (Figure 3.11). The limited set of CHH methylated cytosines contained only statistically identified methylated cytosines from CHH contexts using a binomial distribution (Kawakatsu et al. (2016)). The observation disappeared when all CHH cytosines independent of any site-wise statistical testing were considered (Figure 3.13). Therefore, with CHH methylated cytosines, an increase in CHH methylation at DHSs is observed but overall there is a decrease in CHH methylation. Similarly, Sullivan et al. (2014) noted an increase in CHH methylation in DHSs, although it appears they also used only identified CHH methylated cytosines in their analysis.

Both CHG and CHH methylation are unique to plants and fungi and to date their precise role in any biological function has not been identified (Suzuki and Bird (2008); Lister et al. (2009)). Thus, further investigations into the function of CHG and CHH methylation are needed, particularly in the case of CHH's effect on chromatin structure. Specifically, why does CHH methylation increase at methylated cytosines but decreases across all CHH cytosines? A drop in methylation makes biological sense due to TF binding requiring hypomethylation, but the function of CHH's methylation spike is unknown (Jones et al. (1998); Klose and Bird (2006)). Also of importance to note is the spike in methylation seen for CpG methylation +500 and -500 bp from the centre of DHSs (Figure 3.11). This is believed to be methylation due to TSS effects since the spike disappeared when TSS DHSs were removed (Figure 3.13). Additionally, this is supported as Vining et al. (2012) identified a methylation peak upstream of the TSS. However, in this study the reason a peak is observed +500 and -500 bp, rather than simply upstream, is due to DHSs not being normalized for strand directionality.

Another question is whether hypomethylation develops before chromatin becoming accessible or after? One possible explanation is through observing methylation of highly expressed genes versus repressed genes containing DHSs. Zhang et al. (2012a) suggested poised DHSs, DHSs within non transcribed genes, could become demethylated to become transcriptionally active. However, this remains to be tested for DHSs within this work. In previous research, methylation was found recruiting transcriptional co-repressor molecules that have the potential to alter chromatin structure (Jones et al. (1998); Wade et al. (1999)). As methylation is removed, transcriptional co-repressor molecules are no longer bound and chromatin becomes accessible. A specific example is when methylated cytosines recruit histone deacetylases leading to inaccessible chromatin (Jones et al. (1998)). These findings support methylation arising before chromatin modifications and paints a clear association between chromatin remodelling and DNA methylation. However, Ooi et al. (2007) found *de novo* DNA methylation arising after histone modifications occur. Further, the epigenetic state of genomic regions, i.e. histone modifications, were found to be dependent on the methylated state of that region (Suzuki and Bird (2008); Cedar and Bergman (2009)). Lastly, DNA methyltransferases have been found to specifically target nucleosome bound DNA and may be responsible for such results (Chodavarapu et al. (2010)). Thus, these findings differ and support methylation arising after chromatin modifications.

As there is both evidence for DNA methylation influencing histone modifications and histone modifications influencing DNA methylation, a complex interaction between DNA methylation, chromatin modifications, and chromatin remodelling leading to transcriptional changes likely exists. In sum, there is a clear association and strong connection between DNA methylation, histone modifications, chromatin remodelling, and gene transcription. Future research will be required in order to decipher how they interact and influence one another.

4.4.2 Histone modifications

To further understand the link between DNA methylation, histone modifications and chromatin remodelling; H3K27me3 and H3K4me3 were mapped around DHSs. Caution must be taken with interpreting results as while the data were obtained from the epidermal cell layer, a different promoter was used in the analysis. Results showed unique histone modification patterns around DHSs. H3K27me3 increased 1000 bp upstream and downstream of DHSs with a large decrease within DHSs. H3K4me3 increased 1000 bp upstream and downstream of DHSs with a large increase within DHSs (Figure 3.15). Similar histone patterns were observed in existing DHS datasets supporting the current results (Vierstra et al. (2014); Zhang et al. (2012a); Thurman et al. (2012)). Indeed, previous literature displays quite similar results in terms of the defined histone peaks surrounding DHSs (Figure 3.15; Zhang et al. (2012a)).

Biologically, the observed histone modification patterns within DHSs make sense as H3K27 methylation is linked to silent chromatin and transcription repression while H3K4 methylation is linked to active chromatin and high transcription (Kouzarides (2007); Zhang et al. (2007); Ernst et al. (2011); Thurman et al. (2012); Bernstein et al. (2002)). Thus, taken together with the previous results, DHSs are associated with low DNA methylation, low H3K27me3, and high H3K4me3. All of these are indicative of high TF binding and active transcription as confirmed by results in Figure 3.7. H3K27me3 is particularly important as it is controlled through the Polycomb Group protein complexes. Polycomb Group protein complexes are found critical in gene repression in a tissue and cell-type specific fashion (Zhang et al. (2007)).

As with the methylation analysis, a TSS bias may arise within the observed results due to the high frequency of DHSs around the TSS. Thus, all analyses were repeated with datasets lacking TSS DHSs. For histone modifications, the previous observed results are significantly altered. For instance, while H3K27me3 still increased 1000 bp upstream and downstream of DHSs with a decrease within DHSs, the pattern slightly changed. The increase downstream of DHSs is larger than the increase upstream. Previous literature has shown similar results displaying H3K27me3 increasing around DHSs (Zhang et al. (2012a)). H3K4me3 on the other

hand, now displays a large decrease within DHSs. Zhang et al. (2012a) found very similar H3K4me3 results when mapping histone modifications around DHSs. These results are not unexpected as histone modifications, like H3K4me3, are closely tied to TSSs (Li et al. (2012)). In fact, previous research identified enriched H3K4me3 within promoter DHSs (Vierstra et al. (2014); Thurman et al. (2012); Li et al. (2012)). Additionally, Vierstra et al. (2014) identified H3K4me3 and DNase I sensitivity tightly follow each other and are significantly linked. However, they focused on TSS DHSs as well. Hence, these results reveal histone modifications may not necessarily be linked directly to all DHSs, but may be specific to DHSs within certain genomic locations. Further research is required to identify whether these histone modifications are specific to promoter regions or to promoter DHSs. The drop in H3K27me3 and H3K4me3 within DHSs is not unexpected as DHSs are regions lacking nucleosomes (Albert et al. (2007); Schones et al. (2008)).

The peaks of DNA methylation and histone methylation around DHSs are due to the presence of flanking nucleosomes around DHSs (Thurman et al. (2012); Zhang et al. (2012a)). Results showing similar peaks in histone modifications around DHSs have been previously identified (Vierstra et al. (2014); Thurman et al. (2012); Zhang et al. (2012a)). As mentioned previously, DHSs are also called nucleosome free regions (NFRs) and are found highly localized around TSSs (Albert et al. (2007); Schones et al. (2008)). This is supported by this work's results identifying an enrichment of DHSs centred on the TSS. These NFRs are flanked by highly positioned nucleosomes or strongly phased nucleosome arrays (Wu et al. (2014)). These highly positioned nucleosomes are the first nucleosomes upstream and downstream of the TSS and are labelled as the +1 and -1 nucleosomes (Mavrich et al. (2008); Radman-Livaja and Rando (2010); Mavrich et al. (2008); Wu et al. (2014)). These nucleosomes are important as they acquire histone modifications that contribute to their stability to serve as barriers for nucleosome packing (Mavrich et al. (2008)). Considering H3K27me3, a marker of closed chromatin, is enriched at these nucleosomes supports this theory (Figure 3.15). As a result, the highly positioned nucleosomes surrounding DHSs gather specific histone modifi-

cations or DNA methylation patterns resulting in the observed distinct histone modification peaks. Indeed, the highly positioned nucleosomes are associated with histone variants and histone modifications that may facilitate their stability around DHSs or the eviction of nucleosomes within DHSs (Jiang and Pugh (2009); Kouzarides (2007)). For instance, H3K4me3 was previously found highly enriched at highly positioned nucleosomes around NFRs (Vierstra et al. (2014); Thurman et al. (2012); Zhang et al. (2012a)). Additionally, DNA methylation is linked to nucleosome positioning with DNA methyltransferases found to specifically target nucleosome bound DNA (Chodavarapu et al. (2010)). The peaks in DNA methylation may be explained through the targeting of these highly positioned nucleosomes. Overall, DHSs are flanked by highly positioned nucleosomes which gather histone modifications like H3K27me3 and H3K4me3.

4.5 DHSs and the role of motifs in epidermal and endodermal DE/DA genes

A list of differentially-expressed and simultaneously differentially-accessible (DE/DA) genes were selected to identify potential enriched motifs within gene promoters. It was hoped to identify motifs and combinations of motifs responsible for chromatin remodelling or a cell-type specific response and to associate these motifs with DHSs. Epidermal DE/DA promoters were enriched in motifs belonging to the transcription factor families of AT-hook, Dof, homeodomain, and transcription binding proteins (TBP). In contrast, endodermal DE/DA promoters were enriched in motifs belonging to the transcription factor AT-hook and homeodomain families.

Interestingly, the AT-hook motif is implicated in chromatin remodelling across different organisms (Yun et al. (2012); Aravind and Landsman (1998)). Specifically, the HMG family of proteins significantly effects the structure of chromatin and importantly binds to AT-hook motifs as they contain an AT-hook binding peptide (Thanos and Maniatis (1992); Reeves and

Nissen (1990)). The observation that these DE/DA genes contain binding sites for AT-hook binding proteins which modify chromatin, like HMG proteins, provides evidence as to how they may become differentially-accessible. However, to confirm this, future research into modifying these cis-elements and confirming chromatin structural changes would be needed.

Plotting known motifs across promoter regions in conjunction with DHSs returned a great deal of information. For instance, by knowing what motifs are common across a group of differentially-expressed genes allows one to infer a potential common control mechanisms. In this work, epidermal cell-type promoters were enriched in AT-hook, Dof, homeodomain, and TBPs binding domains. Further testing will confirm if this represents a governing control mechanism for epidermal cell-type expression.

Plotting DHSs allowed for the identification of motifs falling within accessible chromatin. As mentioned previously, for most transcription factors, nucleosomes inhibit TF ability to bind to their respective binding sites (Felsenfeld (1992); Workman and Kingston (1998); Yuan et al. (2005)). In fact, the majority of transcription factor bound motifs are associated within nucleosome free regions (Yuan et al. (2005)). As a result, it is expected that the majority of functional motifs in this work will be located within DHSs. Motifs have pervasive mapping across the genome and sifting out the functional motifs from unfunctional motifs represents a significant challenge. Indeed, motif mapping across a genome always results in extensive mapping at numerous locations across the genome. It is hoped by combining motif data with chromatin accessibility data, one can discern functional motif sites over background noise.

This has great implications when identifying potential targets for future experiments and biotechnology applications. It will aid in weeding out potential motif targets driving gene expression for use in deletion or mutagenesis experiments. At present, these results provide a map of motifs and DHSs of which one may select potential targets for future experiments. Testing these motifs through deletion experiments to identify if they control cell-type specific expression or chromatin alterations will be performed in the future.

One goal of these experiments is to identify the motifs responsible for cell-type expres-

sion and to integrate those results with DHS data to locate regulatory modules causing specific expression profiles. It would be ideal to design synthetic promoters with specific custom expression profiles as opposed to recycling native promoters, as is commonly employed. One could potentially design promoters with certain motifs and genome accessibility in order for it to be expressed in certain cell-types and under certain conditions. By combining accessibility control with motif patterns specific to certain cell-types and conditions, a greater control over cell-type specific expression would be obtained. By having this control over cell-type specific expression, genes advantageous for crop growth or development, but not ideal for human consumption, can have their expression fine tuned within certain cell-types and under certain conditions. For example one could test if the combination of motifs enriched within the epidermal DE/DA genes could be used to create an epidermal specific expression promoter within an accessible region. It is hoped that this data will enable the creation of novel promoters with the ability to fine tune gene expression.

4.6 DHSs and the cold regulated pathway

An important goal of this thesis was to identify chromatin alterations present during the cold acclimation response. Cold and other abiotic stresses have been quite extensively implicated in inducing epigenetic and chromatin alterations (Chinnusamy and Zhu (2009); Luo et al. (2012); Hu et al. (2011)). More specifically, cold itself is linked to huge changes in chromatin state within cold-regulated (*COR*) genes (Mayer et al. (2015)). This study set out to identify these chromatin alterations by identifying the changes in DHSs across the epidermal and endodermal cell layers of the Arabidopsis root during cold acclimation.

4.6.1 Extensive chromatin alterations in response to cold

Previous plant cold acclimation and stress experiments have revealed quite differing results with either more or less down regulated genes compared to upregulated genes (Kreps et al.

(2002); Seki et al. (2002); Lee et al. (2005); Hannah et al. (2005)). Specific cold acclimation expression data produced from 7 days of acclimation revealed more genes downregulated than upregulated (Oono et al. (2006)). Furthermore, another study identified 672 long term up-regulated genes and 915 down regulated genes in response to cold acclimation (Hannah et al. (2005)). Similar to these results, this study revealed a clear decrease in the number of accessible sites due to cold acclimation (Table 3.5). Thus, the extensive downregulation of genes occurs alongside a reduction in accessible chromatin, possibly indicating an association. A recent study looking at DHSs due to dark exposure identified a decrease in the number of DHSs alongside a reduction in gene expression (Liu et al. (2017)).

However, while the number of DHSs decreased overall, the number of genes with an upstream 1000 bp DHS did not decrease significantly (Table 3.5). Therefore, an identical number of genes were accessible but their enhancers may possibly have become inaccessible. These enhancers may be located in other genomic DHSs or within the upstream 1000 bp region if a gene contained two DHSs within their promoter. It would be interesting to identify the genes with two promoter DHSs and identify which genes closed only one. This overall decrease in DHSs may provide a link to significant downregulation of genes due to cold acclimation (Hannah et al. (2005)). Thus, it appears the control of cold acclimation gene expression is controlled by enhancer DHSs closing rather than promoters closing completely.

While distinct alterations in chromatin were found between cell-types and between control-cold datasets, DHS data displayed a high degree of overlap. Indeed, results identified 9,000-11,000 shared DHSs between cell-type and control-cold data. As mentioned, more DHSs were found closed than open in response to cold acclimation. Specifically, 8487 DHSs closed in the epidermis under cold acclimation and 6173 DHSs opened. This corresponded to 1897 genes opening and 3776 genes closing due to cold acclimation within the epidermal cell-type. Thus, more genes under cold acclimation became inactivated or reduced their transcription. Cold acclimation induces extensive changes to chromatin with the main result being more closed chromatin. Unfortunately, no other paper has done this form of analysis with their data and

compared open DHSs versus closed DHSs. Hence, any comparisons with numbers generated from this dataset to others can not be made.

While the number of DHSs decreased due to cold acclimation, the number of DHSs within exons drastically increased (Table 3.5; Figure 3.2). A similar effect has been observed with plants exposed to dark conditions or heat stress (Liu et al. (2017); Sullivan et al. (2014)). However, a precise explanation as to why this increase occurs remains unknown. One explanation identified a link between exonic DHSs and cotranscriptional splicing with exonic DHSs colocalizing with gene promoters (Mercer et al. (2013)). It is possible exonic DHSs may fold into close proximity with the promoter region and bind TFs allowing high transcription. A future direction would be to identify these exonic DHSs and attempt to identify why they are opening under cold or dark acclimation and identify what motifs they contain.

It should be noted that many of these exonic DHSs could in fact be intronic DHSs due to the order of importance for classifying DHSs into genomic locations (see Methods). The DHSs within exonic regions may overlap intronic regions. DHSs opening up in intronic DHSs could be enhancer elements increasing the expression of cold genes. This is important as introns have been shown to frequently contain enhancer binding sites that significantly affect gene expression levels in many different species (Duncker et al. (1997); Lu and Cullen (2003); Rose et al. (2016); Rose (2002)). Thus, it could be DHSs within exonic or intronic regions affecting TF binding and transcriptional output.

Out of 672 upregulated genes more than half were found with an upstream 1000 bp DHS in both control and cold conditions. Results also revealed several cold responsive genes becoming accessible in response to cold. However, several cold responsive genes also became inaccessible, at least in the upstream 1000 bp. Although cold responsive genes closing in response to cold is counter-intuitive it may be due to unannotated TSSs, other genomic annotation errors, or false positives/false negatives (Boyle et al. (2008a)). A more intriguing explanation may be the inactivation of repressor regions allowing transcription of cold responsive genes. Nevertheless, the overall result identified is that the majority of cold responsive genes are open in both

control and cold conditions.

This observation is important when considering the cold response. For any abiotic stress response, rapid stress gene expression will ensure damage is kept to a minimum. This is important when considering that changing a chromatin state is time consuming and very slow (Raser and O'Shea (2004)). It has been previously identified as the rate limiting step in the overall transcriptional response (Barbaric et al. (2001)). In comparison, TF activation and binding occurs quickly. Therefore, having these cold responsive genes highly accessible under control conditions will allow the plant to quickly respond to cold.

Having genes accessible to quickly respond to stress is particularly important when considering the upstream genes of a pathway. These are often the first line genes activated to become highly transcribed under conditions. For instance, the main pathway initiated under cold stress and acclimation is the CBF/DREB pathway (Medina et al. (2011); Thomashow (2010); Lee and Thomashow (2012)). The majority of these genes were found highly accessible in all cell-types and conditions supporting the hypothesis that they are accessible for a quick response (Figure 3.21). The observation that the three *CBF* genes of the CBF/DREB pathway are induced within 15 minutes of cold exposure supports these findings (Gilmour et al. (1998); Shinwari et al. (1998)).

Further, while these upstream cold pathway genes may be highly accessible in both control and cold conditions they are not necessarily highly transcribed under control conditions (Zarka et al. (2003)). As stated previously, DNA methylation, histone modifications, miRNA, TF activation and binding, all influence transcriptional activation (Jones et al. (1998); Klose and Bird (2006); Kouzarides (2007); Zhang et al. (2007); Liu et al. (2017)). A gene may be highly accessible but not highly transcribed due to its binding TF not being activated. As TF activation and binding is quicker than chromatin modifications, a gene that is highly accessible but requiring TF binding will respond quicker. These 672 cold responsive genes could be under control of complex transcriptional factor binding and regulation instead. Additionally, these genes may have a highly accessible upstream region to allow basal expression but

require further open chromatin to initiate high transcription. Specifically, in cold acclimation these shared genes may contain a DHS in the upstream promoter allowing basal expression. However, upon cold exposure exonic or intronic DHSs open which allow further TF binding thereby increasing transcription. For genes that become accessible under cold conditions, a more long term cold response not required during early exposure to cold may be the case.

A majority, 62.1% of epidermal cold DHSs were shared in the epidermal control dataset and 59.5% of endodermal cold DHSs were shared in the endodermal control dataset. Thus, while most cold genes are associated with a DHS in both control and cold conditions, there are still vast chromatin alterations taking place within the genome specific to these cell-types. Whether or not these chromatin changes lead to transcriptional changes or not remains to be identified. Considering roughly 50 of the 672 cold responsive genes were found DE/DA, it seems chromatin alterations influence transcriptional changes under acclimation conditions only slightly. However, DE/DA genes were limited to genes with an upstream 1000 bp DHSs and many other genes may be DE/DA classified due to DHSs in other genomic regions (e.g. exonic/intronic). The DE/DA analysis simply identified 50 genes changing expression due to changes in upstream 1000 bp DHSs. Further testing to identify if changes in DHSs across other genomic locations affect transcription are required. Indeed, gene transcription may be controlled through DHSs in other genomic locations separate from upstream 1000 bp DHSs.

It could be that chromatin alterations influence where and when genes get expressed more than how much or how little. As mentioned previously, the splicing of a gene may be altered by exonic DHSs affecting its function thereby allowing cold acclimation (Mercer et al. (2013)). Chromatin alterations in other genomic locations could lead to changes in transcriptional output that may not be linked to those transcriptional changes. For instance, nucleosome position can allow two binding sites to become spatially connected allowing transcription factor binding and interaction between those sites to initiate transcription (Stünkel et al. (1997)). Although the majority of cold acclimated genes contain a DHS in control conditions, it does not mean their chromatin state is static. Accessibility is not simply an on or off switch, or an open or

closed switch, but a spectrum from completely closed to completely open and everything in between (Zhang et al. (2012a); Liu et al. (2017); He et al. (2012)). Hence, while a site may be accessible within a control and cold condition, it may be more accessible under the cold conditions. However, this needs further investigation. An increase in accessibility at a DHS would enhance TF binding thereby increasing transcription of that gene (Figure 3.10, Boyle et al. (2008a); Zhang et al. (2012a)). One way to test this would be comparing the likelihood ratio between shared DHSs in cold acclimated and control conditions.

In summary, drastic changes to chromatin structure were identified in the Arabidopsis root epidermis and endodermis under cold acclimation. This work focused on identifying differences primarily between DHSs within the upstream 1000 bp region. Results identified that while the total number of DHSs decreased within the upstream 1000 bp, the number of genes with an upstream 1000 bp DHS did not decrease to the same degree. Additionally, more than half of the 672 upregulated cold genes were shared between cold and control conditions. Connections between upstream 1000 bp DHSs and cold acclimation transcriptional changes could not be made in this work (Hannah et al. (2005)). Overall, transcriptional control of stress responsive genes appears to be controlled by individual enhancer DHSs closing rather than entire promoters closing. Indeed, exonic or intronic DHSs drastically increased in the epidermis and endodermis under cold acclimation. Hence, it appears chromatin alterations within other genomic regions, like the exon or intron, may influence total transcriptional output.

4.6.2 Unique motifs enriched within cold accessible genes

Due to the fact the majority of downstream and upstream cold responsive genes were accessible in both control and cold conditions, they must be highly controlled through other factors such as transcription factor activation and binding. Of course, another possible explanation for transcriptional changes may be DHSs opening within other genomic locations. However, this analysis focused on TF binding and activation. In order to get a greater understanding of how the downstream and upstream cold pathway genes are controlled, TF motifs were tested for

enrichment and mapped within the downstream and upstream cold pathway genes. The goal was to identify combinations of motifs responsible for an upstream cold response.

The AP2 transcription factor binding sites enriched within the upstream genes are important when considering the cold transcription factors CBF1, CBF2, and CBF3. CBF proteins belong to the AP2 family of transcriptional activators (Stockinger et al. (1997); Fowler and Thomashow (2002)). Thus, it is not surprising their binding elements are enriched within the set of upstream cold responsive genes (Figure 3.21). It is interesting to note that downstream genes of the three CBF proteins were also enriched for AP2 motifs (Figure 3.19, Figure 3.20). In fact, all of the downstream genes within Figure 3.21 contain at least one AP2 binding site. Additionally, *HOS1* contained three AP2 binding sites possibly indicating a feedback mechanism where CBF proteins regulate HOS1 expression (Figure 3.21). This requires further testing to confirm as no study has looked into this. All of these AP2 binding sites are within DHSs supporting a high potential for function. The specific AP2 motif enriched within this dataset is called the C-repeat/dehydration-responsive (CRT/DRE) regulatory element with the core sequence of CCGAC (Baker et al. (1994); Yamaguchi-Shinozaki and Shinozaki (1994)). This core motif is important as it is required for inducing the expression of many cold responsive genes and is the specific cis-binding site for the *CBF* genes (Baker et al. (1994); Jiang et al. (1996)). Furthermore, this core motif is found in many drought responsive genes as there is a slight overlap between cold induced expression and drought induced expression (Seki et al. (2002)). A significant portion of cold induced damage is more so the result of cold induced dehydration (Steponkus et al. (1998); Steponkus and Webb (1992)).

The bHLH is another interesting TF family found with pervasive motif mapping across this set of upstream genes. It is interesting due to the fact that ICE1, an upstream cold responsive gene, is a bHLH transcriptional activator (Chinnusamy et al. (2003)). As shown in Figure 3.21, ICE1 affects the expression of CBF1, CBF2, and CBF3. However, it affects CBF1 and CBF2 very slightly and not as drastically as it affects CBF3 expression (Chinnusamy et al. (2003)). Thus, the enriched bHLH binding motifs across this upstream pathway make biological sense

as ICE1 binds to this respective motif family. The main motif found mapping within this set of upstream genes has the consensus sequence of a Myc binding site CANNTG, with a significant portion having the G box motif CACGTG (Jakoby et al. (2002); Bailey et al. (2003)). Interestingly, the bHLH transcription factor family binds to DNA with that Myc binding consensus sequence (Meshi and Iwabuchi (1995)). Thus, the bHLH binding domains within the *CBF* DHSs provide support for this pathway's control mechanisms.

In addition to the CBF pathway controlling cold responsive genes, many cold induced genes are controlled through the ABA-dependent pathway (Lang et al. (1994); Laang and Palva (1992); Gilmour et al. (1998)). Induction of these cold responsive genes due to the ABA-dependent pathway is thought to be caused by the interaction of bZIP transcription factors. This is interesting as motifs within the bZIP transcription factor family were enriched in this upstream cold pathway (Figure 3.21). Hence, these sites within DHSs provide a link to the CBF pathway and the ABA independent pathway for cold acclimation.

CG-1 is another interesting transcription factor family found to previously map in the *CBF2* promoter (Finkler et al. (2007); Doherty et al. (2009)). In this work, CG-1 binding motifs were enriched across the upstream pathway, particularly in all three *CBF* genes and *ZAT12*. Additionally, these binding sites were found within the DHSs of the *CBF* and *ZAT12* promoters. These binding sites are critical as they are binding sites for calmodulin transcriptional activators. Upon cold exposure there is a large influx of calcium interacting with the calmodulin transcriptional activators allowing them to bind their CG-1 sites (Doherty et al. (2009); Knight et al. (1991, 1996)). In fact, a mutant copy of *camta3*, one of the calmodulin activators, reduces transcription of CBF1, CBF2, and ZAT12 (Doherty et al. (2009)). This provides a link to the CG-1 binding sites within DHSs and a cold specific response due to calcium induction.

The final transcription factor family motif enriched within this upstream pathway is the MYB-SANT family. The three CBF proteins each carry a MYB-SANT binding motif and all CBF proteins were previously found to interact with MYB15 (Agarwal et al. (2006)). Under normal conditions, MYB15 negatively regulates *CBF* expression by binding to its binding

sites within the three *CBF* genes. Under cold stress MYB15 interacts with ICE1 allowing ICE1 to bind to and activate the expression of the *CBF* genes while preventing MYB15 from inhibiting the *CBF* genes (Zhou et al. (2011); Agarwal et al. (2006)). Therefore, MYB15 acts to inhibit *CBF* expression and enable activation of *CBF* expression. This may be why early research into MYB15 showed confusion as to whether it activated or repressed genes in the cold pathway (Agarwal et al. (2006)). Importantly, under cold stress, the binding sites for MYB15 within *CBF3* are located within closed chromatin, rather than open chromatin. The interaction between MYB15 and ICE1 is expected as research previously found bHLH transcription factors, like ICE1, require co-transcription factors like MYB15 in order to activate target genes (Chinnusamy et al. (2003); Spelt et al. (2000)).

Downstream mapping results shared similar motif enrichment patterns as the upstream pathway, with enrichment of AP2, bZIP, and MYB-SANT TF binding domains. In the context of cold stress and acclimation, such transcription factor binding sites make biological sense. Particularly, for CBF proteins to bind to cold responsive genes, as shown in Figure 3.21, they require the enrichment of AP2 binding sites. Interestingly, the CCGAC AP2 motif was enriched within downstream genes, connecting the downstream cold genes with CBF interaction (Figure 3.19, Figure 3.20).

The MADS box transcription factor family displayed significant enrichment within the epidermal cold DE/DA gene list (Pajoro et al. (2014)). MADS box transcription factors precede changes in chromatin accessibility and have the ability to bind to closed chromatin (Pajoro et al. (2014)). Additionally, MADS box transcription factors bind with nucleosome remodelers or histone modification enzymes, suggesting they may function as pioneering factors affecting the chromatin state (Zhang et al. (2012b); Smaczniak et al. (2012); Pajoro et al. (2014)). Thus, it appears the epidermal cold DE/DA gene list may undergo chromatin alterations due to cold acclimation through the binding of MADS box transcription factor families. Endodermal specific genes were not enriched with motifs belonging to TF families with strong ties to any chromatin remodelling or histone modification potential.

4.6.3 Motif locations for future modifications

One goal of mapping motifs and DHSs within this upstream cold pathway and in the downstream genes was to identify target motifs and/or genes for future cold-resistant crop engineering. Additionally, with DHS integration and analysis it is hoped that a greater control over the chromatin accessibility of transgenes may be obtained. An ideal outcome would be to create cold stress resistant plants without having the plants become cold acclimated. This would allow plants to respond quickly to cold stress and increase plant survivability. This could involve inserting motifs into a promoter to enable activation of select genes or modifying existing motifs so they become inactivated. Having DHS data in combination with motif data makes it significantly easier to identify the functional motif regions from the non-functional motifs for targeted mutagenesis.

Possible sites for modification within this pathway could be the MYB-SANT binding sites in the three *CBF* genes. Deactivating or removing these motifs may potentially alter how they are repressed, allowing these genes to be activated in a more consistent manner. Additionally, MADS domain transcription factor binding domains could be modified to test for a cold resistant plant. This may allow the chromatin state of cold resistant genes to be constantly accessible allowing active and quick transcription. Future experiments can use this available information to target potential motif sites within DHSs for CRISPR/Cas9 modification experiments to obtain a specific transcriptional outcome. Another possible outcome for this research would be the creation of synthetic promoters discussed in the previous section. One could design a promoter for a cold resistance gene with a specific expression profile by taking advantage of accessible chromatin. Furthermore, if a gene is found to be detrimental to the plant in certain cell-types, its expression could be fine tuned with chromatin accessibility or motif placement. In sum, this research has far reaching applications in the design of synthetic genes and promoters by taking advantage of motif and DHS placement.

Chapter 5

Conclusions and future perspectives

5.1 Arabidopsis DNase hypersensitive site importance in cell-type identity and stress response

DNase-seq was utilized within this thesis to identify DHSs within the Arabidopsis epidermal and endodermal root cell-types under control and cold conditions. However, existing DNase-seq protocols could not enable the identification of DHSs from a single cell-type from the Arabidopsis root. Therefore, a novel wet-lab DNase-seq protocol within the Austin lab, deemed DDTS, was used identify DHSs. However, this generated a new form of data with no existing analysis procedures to identify the DHSs. As a result, this work developed a novel bioinformatic tool, called *DDTS*, to identify DHSs from the newly generated data. This tool works by comparing an undigested sample to a digested sample and identifying regions with significant lack of sequencing information in the digested sample compared to the undigested sample. The bioinformatic tool is an easy to use program written general enough for use on a wide range of organisms across various different experimental forms of data. It could potentially be used on data other than DNase-seq generated data like ChIP-seq data. *DDTS* has significant advantages over other DNase-seq wet-lab and analysis procedures in that it is easy to perform on single cell-types, requires minimal equipment, a minimal amount of nuclei/DNA, and minimal

sequencing information. Additionally, new quality control checks were developed to ensure proper digestion of DNA samples and accurate identification of DHSs through *DDTS*.

Thousands of DHSs were identified across the epidermal and endodermal root cell-types across control and cold conditions through the use of *DDTS*. These DHS datasets contained many shared but many unique DHSs indicating a pervasive role of chromatin alterations in cell-type identity and under acclimation conditions. The majority, >60%, of DHSs were shared between the epidermal and endodermal dataset. The remaining 40% percent of DHSs is not unexpected as cell-types within the Arabidopsis root serve highly specific functions. DHSs across all conditions were heavily enriched within the upstream 1000 bp, particularly around the TSS. DHSs showed unique size distributions across genomic locations, with the largest sites located within the upstream 1000 bp. Results indicated that DHSs largely regulate transcriptional output by affecting TF binding within promoter regions of genes. Indeed, the presence of DHSs largely influenced gene transcription with the highest expressed genes more likely associated with DHSs than the lowest transcribed genes. This effect was largely seen for DHSs in the upstream 1000 bp and in the 5'UTR. Furthermore, in addition to the presence of DHSs, the accessibility of DHSs significantly affected gene transcription. Results identified that the most accessible DHSs led to genes with the highest expression. Additionally, the highest accessible DHSs within the 5'UTR and intron had the highest overall expression. This indicates the degree of accessibility significantly affects the ability of TFs to bind to their respective sequences thereby influencing gene expression. In sum, the presence and accessibility of DHSs, mainly in the upstream 1000 bp, significantly affected TF binding thereby influencing gene expression.

To characterize chromatin accessibility across the genome, DHSs were mapped with existing epigenetic data to look for unique epigenetic patterns within and around DHSs. DHSs displayed unique epigenetic patterns within and around them. CpG and CHG methylation were reduced within DHSs across all analyses. CHH methylation, when considering only methylated cytosines, uniquely increased within DHSs. Within the context of all cytosines, CHH methylation dropped within DHSs but increased around DHSs. H3K27me3, an inactivation marker,

was reduced within DHSs while H3K4me3, an activation marker, was enriched in DHSs. However, the enrichment of H3K4me3 was only found located within TSS DHSs. The occurrence of DNA methylation and histone modification peaks around DHSs is explained by highly positioned nucleosomes surrounding DHSs (Jiang and Pugh (2009); Kouzarides (2007)).

The majority of downstream cold acclimation genes contained a DHS in their upstream 1000 bp in both control and cold conditions. Similarly, the majority of upstream cold CBF pathway genes contained a DHS in their upstream 1000 bp under control and cold conditions. It is hypothesized plants have stress responsive genes accessible in all conditions so the plant may respond quickly. Furthermore, while the number of DHSs reduced due to cold, the number of genes with an upstream 1000 bp DHS did not significantly change. Additionally, only 50 DE/DA genes were identified as this work only considered upstream 1000 bp in the analysis. Hence, it is hypothesized that upstream 1000 bp DHSs do not significantly alter transcription due to cold acclimation. In support of this, a large increase in exonic or intronic DHSs under cold acclimation was identified. A hypothesis is that exonic or intronic DHSs become accessible due to cold acclimation, thus affecting transcription of cold responsive genes via TF binding sites in exonic or intronic regions. However, future work to support this is needed.

Furthermore, *a priori* scanning of existing TF binding motifs in select DE/DA genes of cell-type specific and upstream cold induced genes identified unique results. This was performed in hope of identifying motifs responsible for chromatin changes seen within these genes. The epidermal and endodermal cell-types shared enriched binding domains for the TF family of AThook. Interestingly, the AThook TF family has strong ties to chromatin remodelling. Cold DE/DA genes were enriched with many AP2 binding sites, specifically CCGAC, a known cold binding domain. Additionally, MADS box binding domains were enriched within these cold DE/DA genes pointing to a possible mechanism as to how they become accessible under cold conditions. Lastly, the upstream cold CBF pathway was enriched in binding domains for the transcription factors AP2, bZIP, Myb-SANT, TBP, bHLH, CG-1, and homeodomain.

Ultimately, through mapping motif data with DHS data it is hoped the identification of

functional motifs will be made easier. Functional motifs are those with ability to bind TFs since their regions are highly accessible. Results showed DHSs heavily enriched in motifs for various TF families. Integrating epidermal and endodermal motif data with DHS data will enable the identification of functional motifs that lead to cell identity or control cell-type specific chromatin accessibility. Cold motif data and DHS data will enable the identification of motifs responsible for cold response or for altering chromatin accessibility under cold acclimation. The ultimate goal being to identify motifs responsible for cell identity, chromatin alterations, and stress response within DHSs. This has great promise identifying potential targets for future experiments and biotechnology applications. Mutagenesis or CRISPR/Cas9 experiments benefit from this research by narrowing down the motifs to select for experimentation. Additionally, it is hoped this research will enable the creation of a motif ”tool-box” to control cell-type specific and abiotic stress expression. This ”tool-box” would not only contain the TF binding motifs for cell-type specific and stress response expression but also the motifs or processes to control where and when that motif will be accessible. Controlling not only the position of motifs in a promoter but also when that promoter will become accessible will enable a greater control over transgene expression.

In conclusion, DHSs displayed unique and shared characteristics across cell-types and abiotic stress conditions. Results found DHSs change drastically across cell-types and under cold acclimation. Finally chromatin accessibility displayed transcriptional control, has defined epigenetic characteristics, and has the potential to be used in conjunction with motif data for biotechnology purposes.

5.2 Study limitations

DHSs were successfully identified and integrated with various data sources across the Arabidopsis root epidermis and endodermis under control and cold conditions through the use of a novel protocol and analysis procedure called DDTS. However, the identification of DHSs and

integration with other data sources came with several limitations. There are significant challenges facing not only DDTS but other DNase-seq protocols in generating and processing the raw data.

While *DDTS* was successful at identifying DHSs with limited sequencing information, *DDTS* is highly limited by the amount of sequencing information used for analysis. As sequencing reads across datasets were downsampled, the quality of DHSs degraded and the identification of DHSs was significantly more difficult (results not shown). Specifically, the false positive rate increased significantly with a lower read count and DHS width increased. Therefore, it is necessary to obtain a high amount of sequencing information in order to produce high quality results. However, sequencing is time consuming and costly, thus a compromise must be made between quality and the amount of dollars spent. The more sequencing data obtained, the more power an experiment will have in identifying high quality DHSs. It is suggested that increasing the number of reads above what was used within this thesis will enable more power in identifying high quality DHSs and lower the false positive rate. However, 10 million reads should suffice to produce results for the Arabidopsis genome.

Additionally, there are many settings that can be altered within DDTS and helper programs, like F-Seq, that may significantly alter results. Changing any of the countless settings may increase power of DHS identification or decrease the power. Additionally, the false positive rate or the false negative rate could significantly increase or decrease. One may change these settings to increase the effectiveness of identifying DHSs, however, that can lead to complex analysis procedures of constantly changing and testing results. There are many settings and combinations of settings to alter, potentially leading to time consuming analysis procedures. It is best to select settings through an educated guess based on past knowledge and biological information at hand. In sum, *DDTS* worked effectively in identifying thousands of DHSs but caution should be used in analysis procedures.

The use of external transcriptional and epigenetic data produces many analysis and data integration limitations making interpretation of the results difficult. For example, RNA-seq,

Methyl-seq, and ChIP-seq contain many challenges in processing and analysing their results. For any of these procedures, proper analysis is difficult and wildly different results can be obtained across not only biological replicates but also technical replicates. All of these procedures have differing analysis tools and programs that can produce different results from the same data. Similarly with DDTS, slight changes in settings can produce differing results.

Further, similar growing conditions are a challenge when comparing data from various sources. A change in lighting, water, heat, and soil conditions can significantly affect results and cause plants to display wildly different responses. This work grew plants in liquid media in order to obtain the required number of nuclei for DDTS. Growing Arabidopsis in liquid media is expected to significantly affect transcription profiles compared to growing it in soil. Therefore, comparing data between several sources and interpreting information should be done carefully, especially when looking at fine details. Any generalizations must be made carefully and should be investigated and supported further with additional experimentation. However, when looking at general trends or general observations, comparing data across various sources is a powerful tool. It enables distinct observations to be made in a cost and time efficient manner. Any conclusions made should be supported with previous literature or with further experimentation.

An additional difficulty with the external data sources used in this work concerned each experiment's procedure to isolate cell-type specific nuclei. While Li et al. (2016) and Kawakatsu et al. (2016) isolated nuclei using the same Werewolf (wer) promoter, as used in this work, they utilized FACS for nuclei isolation. Despite biological variation being reduced through the use of the same promoter, it is unknown whether FACS and INTACT would produce differing samples of isolated nuclei. Each protocol may produce slightly differing nuclei pools that can significantly affect results. Furthermore, while Deal and Henikoff (2010) used the INTACT protocol to isolate the same cell-type specific nuclei, they did not use the same promoter to isolate nuclei. Their data were used to identify if any trends could be observed within DHSs and specific details could not be pulled out from this analysis. However, data were obtained from

a similar root cell-type using the promoter of *GL2*. Unlike the Werewolf promoter isolating the entire epidermis, the *GL2* promoter isolates only the non-hair cells of the root epidermis. Therefore, any discrepancies should be reduced to a minimum as differences within the same tissue will be kept to a minimum. Additionally, this analysis only observed averaged trends across a group of sites thereby reducing potential errors.

Lastly, genomics generates a vast amount of data that is difficult to interpret and analyze. For example, as of 2016, over 4 x 10^{15} base pairs of sequencing information are available on-line in public repositories (Leinonen et al. (2010); Kanz et al. (2005); Benson et al. (2012); Barrett et al. (2012)). While the issue of computing power and analysis programs is certainly an important one requiring solutions, it is definitely not the main issue. The main issue regarding this vast amount of data is in the man power to analyze and interpret the data in a biological meaningful way. While the cost of sequencing has drastically reduced, the cost of analysis and of human resources has now become the main part of any lab's budget (Sboner et al. (2011)). Cost aside, the human aspect is a significant limiting factor in regards to large experiments. Proper interpretation and making sense of data is a challenging task and often involves many sets of eyes (Mardis (2010)). This difficulty was observed within this work particularly with the integration of the data with other sources. It is often difficult to know how to process gigabytes of data or even what to look for in results. Reducing data to a meaningful result through data reduction is often a difficult task within genomics. Additionally, genomics suffers from analysing with a wide brush rather than having specific scientific questions requiring answers. While often bioinformaticians are needed to analyze this vast amount of data, biologists are also required to interpret it in a meaningful fashion. Work flows need to be made to enable scientists to focus on analysing and interpreting the data in the most efficient manner possible. Within any of these genomic experiments there will always be some result that goes missed.

5.3 Future directions

There are many possible future directions and improvements for the work presented here. A number of these have been discussed in the previous chapters but many additional avenues for future direction are possible. Whenever working with vast amounts of data, there will always be more questions requiring answers than those answered. For instance, there is continued work in analysing differences in accessibility between shared DHSs, predicting new regulatory elements from DHSs, understanding why DHSs display unique epigenetic patterns, exonic or intronic DHS role in cold transcriptional control, and future experiments utilizing DHS data in combination with motif data.

A possible future direction with DHS data is to test how shared DHSs across conditions and cell-types change in accessibility. As mentioned previously, accessibility is not an on or off switch. Rather, it is instead more akin to a dimmer switch from completely closed to completely open and everything in between. While a site may be accessible in both control and cold conditions, it may be more accessible under cold conditions. However, this remains to be tested with the DHSs generated in this work. It is still unknown whether this type of analysis will work with this custom *DDTS* protocol. Comparing the likelihood fold ratio between DHSs across conditions could work in theory, however it is unknown if it is practical to compare across datasets. Additionally, DHSs could be used within Arabidopsis to predict enhancer elements. Enhancer elements are seemingly rare within Arabidopsis. Thus, utilizing DHSs to identify new enhancer elements, shown successful in previous experiments, is a great avenue to continue this work (Jiang (2015)). However, connecting distal elements to their respective genes is a difficult task to overcome.

In addition, while a connection was made between DHSs, methylation, histone modifications, and transcription; how exactly these processes interact and influence one another is largely unknown The order of events or even how one factor influences another has been the focus of many research experiments but its answers still remain unknown. Despite this, many studies have reported the link between these individual factors and displayed their defining pat-

terns (Zhang et al. (2012a); Sullivan et al. (2014); Thurman et al. (2012); Vierstra et al. (2014); Zhang et al. (2012a); Thurman et al. (2012)).

Furthermore, a unique possible avenue to continue research is in the role of exonic or intronic DHSs in cold transcriptional control. Results identified an increase in exonic or intronic DHSs due to cold acclimation in both the epidermis and endodermis. While possible theories exist to explain this observation, no conclusive answer has been found. Thus, future work in identifying exonic or intronic DHS's role in cold acclimation, particularly in a transcriptional response, is required.

Lastly, this type of research may be used in the future for designing synthetic promoters. This work looked into aspects of genomics from motif data, chromatin alterations, histone modifications, methylation, and transcription. It is hoped by analysing all biological aspects affecting transcription, a better promoter can be developed. A promoter designed to take into consideration various aspects of motif placement and combination, individual methylation patterns, histone modifications, and DHSs, could enable a greater control over expression at the cell-type and stress response level. A promoter designed with specific TF binding motifs and DHSs for cell-type specific expression will enable a greater control over its expression level spatially and temporally. Promoters can be intelligently designed to be slightly accessible or highly accessible allowing its expression to be fine tuned. This is why this thesis identified motifs enriched within gene sets and under abiotic stress conditions. Additionally, it is also why this work set out to show a connection between a DHS's accessibility and transcriptional output.

DHS data is a powerful tool for use in mapping *a priori* motif sites and *de novo* motif sites. The discussion mentioned DHSs could be used to identify functional motifs for use in future experiments. While this was not validated here, previous work found strong connections between TF binding and accessible chromatin painting a clear picture of chromatin interfering with TF binding (Zhang et al. (2014); Yuan et al. (2005)). As a result, motifs within DHSs have an increase likelihood of functionality and DHSs can be used to more accurately identify

functional motifs bound to TFs. Lastly, DHS data has shown great promise in enabling more powerful identification of novel binding motifs (Sullivan et al. (2014)). Motif algorithms require a base pair input in order to predict novel motifs and the size of the input sequence can significantly affect results. However, by feeding these algorithms only functional regions, i.e. DHSs, the input sequence is significantly reduced, thereby enabling a greater statistical power in identifying novel motifs.

Complete epigenetic maps for various tissues and cell-types including all aspects of possible data from DNA methylation, to histone modifications, to chromatin accessibility are needed to fully understand epigenetics role in transcription and gene expression. Indeed, epigenetic maps across various tissues, cell-types, and stresses is essential for a future work to completely understand the nuances of gene transcription and regulation. Additionally, further research connecting how chromatin accessibility influences expression under acclimation conditions is required. Trying to decipher why some genes are always accessible against those that become accessible under cold acclimation will be necessary for understanding chromatin's role in acclimation conditions.

Bibliography

Abdelkarim, B. T., Maranda, V., and Drouin, G. (2017). The fate of retrotransposed processed genes in *Arabidopsis thaliana*. *Gene*, 609:1–8.

Adams, C. C. and Workman, J. L. (1995). Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Molecular and Cellular Biology*, 15(3):1405–1421.

Adkins, M. W., Howar, S. R., and Tyler, J. K. (2004). Chromatin disassembly mediated by the histone chaperone Asf1 is essential for transcriptional activation of the yeast PHO5 and PHO8 genes. *Molecular Cell*, 14(5):657–666.

Agarwal, M., Hao, Y., Kapoor, A., Dong, C.-H., Fujii, H., Zheng, X., and Zhu, J.-K. (2006). A R2R3 type MYB transcription factor is involved in the cold regulation of CBF genes and in acquired freezing tolerance. *Journal of Biological Chemistry*, 281(49):37636–37645.

Albert, I., Mavrich, T. N., Tomsho, L. P., Qi, J., Zanton, S. J., Schuster, S. C., and Pugh, B. F. (2007). Translational and rotational settings of H2A. Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature*, 446(7135):572.

Allen, B. L. and Taatjes, D. J. (2015). The Mediator complex: a central integrator of transcription. *Nature Reviews. Molecular cell biology*, 16(3):155.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11:R106.

Aravind, L. and Landsman, D. (1998). AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nuclei Acids Research*, 26(19):4413–4421.

Arnone, M. I. and Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–1864.

Austin, R. S., Hiu, S., Waese, J., Ierullo, M., Pasha, A., Wang, T. T., Fan, J., Foong, C., Breit, R., Desveaux, D., et al. (2016). New BAR tools for mining expression data and exploring Cis-elements in *Arabidopsis thaliana*. *The Plant Journal*, 88(3):490–504.

Bailey, P. C., Martin, C., Toledo-Ortiz, G., Quail, P. H., Huq, E., Heim, M. A., Jakoby, M., Werber, M., and Weisshaar, B. (2003). Update on the basic helix-loop-helix transcription factor gene family in *Arabidopsis thaliana*. *The Plant Cell*, 15(11):2497–2502.

Baker, S. S., Wilhelm, K. S., and Thomashow, M. F. (1994). The 5-region of *Arabidopsis thaliana* cor15a has cis-acting elements that confer cold-, drought-and ABA-regulated gene expression. *Plant Molecular Biology*, 24(5):701–713.

Barbaric, S., Walker, J., Schmid, A., Svejstrup, J., and Hörz, W. (2001). Increasing the rate of chromatin remodeling and gene activation—a novel role for the histone acetyltransferase Gcn5. *The EMBO Journal*, 20(17):4944–4951.

Barlow, D. P. (1993). Methylation and imprinting: from host defense to gene regulation? *Science*, 260(5106):309–311.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2012). Ncbi GEO: archive for functional genomics data setsupdate. *Nuclei Acids Research*, 41(D1):D991–D995.

Becker, P. B. and Hörz, W. (2002). ATP-dependent nucleosome remodeling. *Annual Review of Biochemistry*, 71(1):247–273.

Bell, O., Tiwari, V. K., Thomä, N. H., and Schübeler, D. (2011). Determinants and dynamics of genome accessibility. *Nature Reviews Genetics*, 12(8):554–564.

Benfey, P. N. and Schiefelbein, J. W. (1994). Getting to the root of plant development: the genetics of Arabidopsis root formation. *Trends in Genetics*, 10(3):84–88.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2012). Genbank. *Nuclei Acids Research*, 41(D1):D36–D42.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2015). The Arabidopsis information resource: making and mining the gold standard annotated reference plant genome. *Genesis*, 53(8):474–485.

Bernstein, B. E., Humphrey, E. L., Erlich, R. L., Schneider, R., Bouman, P., Liu, J. S., Kouzarides, T., and Schreiber, S. L. (2002). Methylation of histone H3 Lys 4 in coding regions of active genes. *Proceedings of the National Academy of Sciences*, 99(13):8695–8700.

Bianchi, M., Crinelli, R., Giacomini, E., Carloni, E., and Magnani, M. (2009). A potent enhancer element in the 5-UTR intron is crucial for transcriptional regulation of the human ubiquitin C gene. *Gene*, 448(1):88–101.

Birnbaum, K., Shasha, D. E., Wang, J. Y., Jung, J. W., Lambert, G. M., Galbraith, D. W., and Benfey, P. N. (2003). A gene expression map of the Arabidopsis root. *Science*, 302(5652):1956–1960.

Board, J., Peterson, M., and Ng, E. (1980). Floret sterility in rice in a cool environment. *Agronomy Journal*, 72(3):483–487.

Bondino, H. G. and Valle, E. M. (2009). A small intergenic region drives exclusive tissue-specific expression of the adjacent genes in *Arabidopsis thaliana*. *BMC Molecular Biology*, 10(1):95.

Bonner, W., Hulett, H., Sweet, R., and Herzenberg, L. (1972). Fluorescence activated cell sorting. *Review of Scientific Instruments*, 43(3):404–409.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008a). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322.

Boyle, A. P., Guinney, J., Crawford, G. E., and Furey, T. S. (2008b). F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, 24(21):2537–2538.

Brehm, A., Längst, G., Kehle, J., Clapier, C. R., Imhof, A., Eberharter, A., Müller, J., and Becker, P. B. (2000). dMi-2 and ISWI chromatin remodelling factors have distinct nucleosome binding and mobilization properties. *The EMBO Journal*, 19(16):4332–4341.

Bucceri, A., Kapitza, K., and Thoma, F. (2006). Rapid accessibility of nucleosomal DNA in yeast on a second time scale. *The EMBO Journal*, 25(13):3123–3132.

Bussemaker, H. J., Li, H., and Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, 27(2).

Carlson, M. (2016). *ath1121501.db: Affymetrix Arabidopsis ATH1 Genome Array annotation data (chip ath1121501)*. R package version 3.2.3.

Cedar, H. and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews. Genetics*, 10(5):295.

Chinnusamy, V., Ohta, M., Kanrar, S., Lee, B.-h., Hong, X., Agarwal, M., and Zhu, J.-K. (2003). ICE1: a regulator of cold-induced transcriptome and freezing tolerance in Arabidopsis. *Genes & Development*, 17(8):1043–1054.

Chinnusamy, V., Zhu, J., and Zhu, J.-K. (2007). Cold stress regulation of gene expression in plants. *Trends in Plant Science*, 12(10):444–451.

Chinnusamy, V. and Zhu, J.-K. (2009). Epigenetic regulation of stress responses in plants. *Current Opinion in Plant Biology*, 12(2):133–139.

Chodavarapu, R. K., Feng, S., Bernatavichute, Y. V., Chen, P.-Y., Stroud, H., Yu, Y., Hetzel, J., Kuo, F., Kim, J., Cokus, S. J., et al. (2010). Relationship between nucleosome positioning and DNA methylation. *Nature*, 466(7304):388.

Clapier, C. R. and Cairns, B. R. (2009). The biology of chromatin remodeling complexes. *Annual Review of Biochemistry*, 78:273–304.

Costa, S. and Shaw, P. (2006). Chromatin organization and cell fate switch respond to positional information in Arabidopsis. *Nature*, 439(7075):493.

Crawford, G. E., Davis, S., Scacheri, P. C., Renaud, G., Halawi, M. J., Erdos, M. R., Green, R., Meltzer, P. S., Wolfsberg, T. G., and Collins, F. S. (2006). DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature Methods*, 3(7):503–509.

Cumbie, J. S., Filichkin, S. A., and Megraw, M. (2015). Improved DNase-seq protocol facilitates high resolution mapping of DNase I hypersensitive sites in roots in *Arabidopsis thaliana*. *Plant Methods*, 11(1):42.

Darnell, J. E. (2013). Reflections on the history of pre-mRNA processing and highlights of current knowledge: a unified picture. *RNA*, 19(4):443–460.

Deal, R. B. and Henikoff, S. (2010). A simple method for gene expression and chromatin profiling of individual cell types within a tissue. *Developmental Cell*, 18(6):1030–1040.

Deal, R. B. and Henikoff, S. (2011). The INTACT method for cell type-specific gene expression and chromatin profiling in *Arabidopsis thaliana*. *Nature Protocols*, 6(1):56–68.

Di Laurenzio, L., Wysocka-Diller, J., Malamy, J. E., Pysh, L., Helariutta, Y., Freshour, G., Hahn, M. G., Feldmann, K. A., and Benfey, P. N. (1996). The SCARECROW gene regulates

an asymmetric cell division that is essential for generating the radial organization of the Arabidopsis root. *Cell*, 86(3):423–433.

Dietz, K.-J., Vogel, M. O., and Viehhauser, A. (2010). AP2/EREBP transcription factors are part of gene regulatory networks and integrate metabolic, hormonal and environmental signals in stress acclimation and retrograde signalling. *Protoplasma*, 245(1-4):3–14.

Dinno, A. (2017). *dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums*. R package version 1.3.4.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Doherty, C. J., Van Buskirk, H. A., Myers, S. J., and Thomashow, M. F. (2009). Roles for Arabidopsis CAMTA transcription factors in cold-regulated gene expression and freezing tolerance. *The Plant Cell*, 21(3):972–984.

Dolan, L., Janmaat, K., Willemsen, V., Linstead, P., Poethig, S., Roberts, K., and Scheres, B. (1993). Cellular organisation of the *Arabidopsis thaliana* root. *Development*, 119(1):71–84.

Duncker, B., Davies, P., and Walker, V. (1997). Introns boost transgene expression in *Drosophila melanogaster*. *Molecular and General Genetics MGG*, 254(3):291–296.

Enstone, D. E., Peterson, C. A., and Ma, F. (2002). Root endodermis and exodermis: structure, function, and responses to the environment. *Journal of Plant Growth Regulation*, 21(4):335–351.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Systematic analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43.

Esau, K. (1977). *Anatomy of Seed Plants. 2nd edn. J.* Wiley and Sons, New York.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research*, 8(3):175–185.

Feil, R. and Fraga, M. F. (2012). Epigenetics and the environment: emerging patterns and implications. *Nature Reviews. Genetics*, 13(2):97.

Felsenfeld, G. (1992). Chromatin as an essential part of the transcriptional mechanism. *Nature*, 355(6357):219.

Filichkin, S. A. and Megraw, M. (2017). DNase I SIM: a simplified in-nucleus method for DNase I hypersensitive site seqencing. *Methods in Molecular Biology (Clifton, NJ)*, 1629:141.

Finkler, A., Ashery-Padan, R., and Fromm, H. (2007). CAMTAs: calmodulin-binding transcription activators from plants to human. *Febs Letters*, 581(21):3893–3898.

Fowler, S. and Thomashow, M. F. (2002). Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. *The Plant Cell*, 14(8):1675–1690.

Francis, N. J., Kingston, R. E., and Woodcock, C. L. (2004). Chromatin compaction by a polycomb group protein complex. *Science*, 306(5701):1574–1577.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315.

Gilmour, S. J., Fowler, S. G., and Thomashow, M. F. (2004). Arabidopsis transcriptional activators CBF1, CBF2, and CBF3 have matching functional activities. *Plant Molecular Biology*, 54(5):767–781.

Gilmour, S. J. and Thomashow, M. F. (1991). Cold acclimation and cold-regulated gene expression in ABA mutants of *Arabidopsis thaliana*. *Plant Molecular Biology*, 17(6):1233–1240.

Gilmour, S. J., Zarka, D. G., Stockinger, E. J., Salazar, M. P., Houghton, J. M., and Thomashow, M. F. (1998). Low temperature regulation of the Arabidopsis CBF family of AP2 transcriptional activators as an early step in cold-induced COR gene expression. *The Plant Journal*, 16(4):433–442.

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., and Toulmin, C. (2010). Food security: the challenge of feeding 9 billion people. *Science*, 327(5967):812–818.

Gross, D. S. and Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. *Annual Review of Biochemistry*, 57(1):159–197.

Gruss, P., Lai, C.-J., Dhar, R., and Khoury, G. (1979). Splicing as a requirement for biogenesis of functional 16S mRNA of simian virus 40. *Proceedings of the National Academy of Sciences*, 76(9):4317–4321.

Gu, Z., Eils, R., and Schlesner, M. (2016). HilbertCurve: an R/Bioconductor package for high-resolution visualization of genomic data. *Bioinformatics*, 32(15):2372–2374.

Guertin, M. J. and Lis, J. T. (2013). Mechanisms by which transcription factors gain access to target sequence elements in chromatin. *Current Opinion in Genetics & Development*, 23(2):116–123.

Guy, C. L., Niemi, K. J., and Brambl, R. (1985). Altered gene expression during cold acclimation of spinach. *Proceedings of the National Academy of Sciences*, 82(11):3673–3677.

Hannah, M. A., Heyer, A. G., and Hincha, D. K. (2005). A global survey of gene regulation during cold acclimation in *Arabidopsis thaliana*. *PLoS Genet*, 1(2):e26.

He, H. H., Meyer, C. A., Chen, M. W., Jordan, V. C., Brown, M., and Liu, X. S. (2012). Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Research*, 22(6):1015–1025.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311.

Henikoff, S. (2008). Nucleosome destabilization in the epigenetic regulation of gene expression. *Nature Reviews Genetics*, 9(1):15–26.

Henry, A., Cal, A. J., Batoto, T. C., Torres, R. O., and Serraj, R. (2012). Root attributes affecting water uptake of rice (*Oryza sativa*) under drought. *Journal of Experimental Botany*, 63(13):4751–4763.

Hernandez-Garcia, C. M. and Finer, J. J. (2014). Identification and validation of promoters and cis-acting regulatory elements. *Plant Science*, 217:109–119.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., et al. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, 6(4):283–289.

Howbrook, D. N., van der Valk, A. M., O'Shaughnessy, M. C., Sarker, D. K., Baker, S. C., and Lloyd, A. W. (2003). Developments in microarray technologies. *Drug Discovery Today*, 8(14):642–651.

Hu, Y., Zhang, L., Zhao, L., Li, J., He, S., Zhou, K., Yang, F., Huang, M., Jiang, L., and Li, L. (2011). Trichostatin A selectively suppresses the cold-induced transcription of the ZmDREB1 gene in maize. *PLoS One*, 6(7):e22132.

Hurst, L. D. (2017). It's easier to get along with the quiet neighbours. *Molecular Systems Biology*, 13(9):943.

Jaglo-Ottosen, K. R., Gilmour, S. J., Zarka, D. G., Schabenberger, O., and Thomashow, M. F.

(1998). Arabidopsis CBF1 overexpression induces COR genes and enhances freezing tolerance. *Science*, 280(5360):104–106.

Jakoby, M., Weisshaar, B., Dröge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J., Kroj, T., and Parcy, F. (2002). bZIP transcription factors in Arabidopsis. *Trends in Plant Science*, 7(3):106–111.

Jiang, C., Iu, B., and Singh, J. (1996). Requirement of a CCGAC cis-acting element for cold induction of the BN115 gene from winter *Brassica napus*. *Plant Molecular Biology*, 30(3):679–684.

Jiang, C. and Pugh, B. F. (2009). Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews. Genetics*, 10(3):161.

Jiang, J. (2015). The 'dark matter' in the plant genomes: non-coding and unannotated DNA sequences associated with open chromatin. *Current Opinion in Plant Biology*, 24:17–23.

Jin, W., Tang, Q., Wan, M., Cui, K., Zhang, Y., Ren, G., Ni, B., Sklar, J., Przytycka, T. M., Childs, R., et al. (2015). Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature*, 528(7580):142–146.

John, S., Sabo, P. J., Thurman, R. E., Sung, M.-H., Biddie, S. C., Johnson, T. A., Hager, G. L., and Stamatoyannopoulos, J. A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43(3):264–268.

Jones, P. L., Veenstra, G. C. J., Wade, P. A., Vermaak, D., Kass, S. U., Landsberger, N., Strouboulis, J., and Wolffe, A. P. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genetics*, 19(2):187–191.

Jung, J.-H., Park, J.-H., Lee, S., To, T. K., Kim, J.-M., Seki, M., and Park, C.-M. (2013). The cold signaling attenuator HIGH EXPRESSION OF OSMOTICALLY RESPONSIVE

GENE1 activates FLOWERING LOCUS C transcription via chromatin remodeling under short-term cold stress in Arabidopsis. *The Plant Cell*, 25(11):4378–4390.

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., et al. (2005). The EMBL nucleotide sequence database. *Nuclei Acids Research*, 33(suppl_1):D29–D33.

Kargiotidou, A., Kappas, I., Tsaftaris, A., Galanopoulou, D., and Farmaki, T. (2010). Cold acclimation and low temperature resistance in cotton: *Gossypium hirsutum* phospholipase Dα isoforms are differentially regulated by temperature and light. *Journal of Experimental Botany*, 61(11):2991–3002.

Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, 107(7):2926–2931.

Kass, S. U., Pruss, D., and Wolffe, A. P. (1997). How does DNA methylation repress transcription? *Trends in Genetics*, 13(11):444–449.

Kasuga, M., Liu, Q., Miura, S., Yamaguchi-Shinozaki, K., and Shinozaki, K. (1999). Improving plant drought, salt, and freezing tolerance by gene transfer of a single stress-inducible transcription factor. *Nature Biotechnology*, 17(3):287–291.

Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., Feldblyum, T., Nierman, W., Benito, M. I., Lin, X., et al. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796–815.

Kawakatsu, T., Stuart, T., Valdes, M., Breakfield, N., Schmitz, R. J., Nery, J. R., Urich, M. A., Han, X., Lister, R., Benfey, P. N., et al. (2016). Unique cell-type-specific patterns of DNA methylation in the root meristem. *Nature Plants*, 2(5):16058–16058.

Keene, M. A., Corces, V., Lowenhaupt, K., and Elgin, S. (1981). DNase I hypersensitive sites in Drosophila chromatin occur at the 5'ends of regions of transcription. *Proceedings of the National Academy of Sciences*, 78(1):143–146.

Kiefer, J. C. (2007). Epigenetics in development. *Developmental Dynamics*, 236(4):1144–1156.

Kiegle, E., Moore, C. A., Haseloff, J., Tester, M. A., and Knight, M. R. (2000). Cell-type-specific calcium responses to drought, salt and cold in the Arabidopsis root. *The Plant Journal*, 23(2):267–278.

Kim, H.-J., Hyun, Y., Jin-Young, P., Mi-Jin, P., Park, M.-K., Kim, M. D., Kim, H.-J., Lee, M. H., Moon, J., Lee, I., et al. (2004). A genetic link between cold responses and flowering time through FVE in *Arabidopsis thaliana*. *Nature Genetics*, 36(2):167.

Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880.

Klose, R. J. and Bird, A. P. (2006). Genomic DNA methylation: the mark and its mediators. *Trends in Biochemical Sciences*, 31(2):89–97.

Knight, H., Trewavas, A. J., and Knight, M. R. (1996). Cold calcium signaling in Arabidopsis involves two cellular pools and a change in calcium signature after acclimation. *The Plant Cell*, 8(3):489–503.

Knight, M. R., Campbell, A. K., et al. (1991). Transgenic plant aequorin reports the effects of touch and cold-shock and elicitors on cytoplasmic calcium. *Nature*, 352(6335):524.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128(4):693–705.

Kreps, J. A., Wu, Y., Chang, H.-S., Zhu, T., Wang, X., and Harper, J. F. (2002). Transcriptome

changes for Arabidopsis in response to salt, osmotic, and cold stress. *Plant Physiology*, 130(4):2129–2141.

Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572.

Kwak, K. J., Kim, J. Y., Kim, Y. O., and Kang, H. (2007). Characterization of transgenic Arabidopsis plants overexpressing high mobility group B proteins under high salinity, drought or cold stress. *Plant and Cell Physiology*, 48(2):221–231.

Laang, V. and Palva, E. T. (1992). The expression of a rab-related gene, *rab18*, is induced by abscisic acid during the cold acclimation process of *Arabidopsis thaliana*. *Plant Molecular Biology*, 20(5):951–962.

Lang, V., Mantyla, E., Welin, B., Sundberg, B., and Palva, E. T. (1994). Alterations in water status, endogenous abscisic acid content, and expression of *rab18* gene during the development of freezing tolerance in *Arabidopsis thaliana*. *Plant Physiology*, 104(4):1341–1349.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

Lee, B.-h., Henderson, D. A., and Zhu, J.-K. (2005). The Arabidopsis cold-responsive transcriptome and its regulation by ICE1. *The Plant Cell*, 17(11):3155–3175.

Lee, C.-M. and Thomashow, M. F. (2012). Photoperiodic regulation of the C-repeat binding factor (CBF) cold acclimation pathway and freezing tolerance in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*, 109(37):15054–15059.

Lee, M. M. and Schiefelbein, J. (1999). WEREWOLF, a MYB-related protein in Arabidopsis, is a position-dependent regulator of epidermal cell patterning. *Cell*, 99(5):473–483.

Leinonen, R., Sugawara, H., Shumway, M., and Collaboration, I. N. S. D. (2010). The sequence read archive. *Nuclei Acids Research*, 39(suppl_1):D19–D21.

Lenth, R. V. (2016). Least-squares Means: The R Package lsmeans. *Journal of Statistical Software*, 69(1):1–33.

Levy-Wilson, B., Paulweber, B., Nagy, B. P., Ludwig, E., and Brooks, A. (1992). Nuclease-hypersensitive sites define a region with enhancer activity in the third intron of the human apolipoprotein B gene. *Journal of Biological Chemistry*, 267(26):18735–18743.

Li, G., Chandrasekharan, M. B., Wolffe, A. P., and Hall, T. C. (2001). Chromatin structure and phaseolin gene regulation. *Plant Molecular Biology*, 46(2):121–129.

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1):84–98.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, S., Yamada, M., Han, X., Ohler, U., and Benfey, P. N. (2016). High-resolution expression map of the Arabidopsis root reveals alternative splicing and lincRNA regulation. *Developmental Cell*, 39(4):508–522.

Liao, Y., Smyth, G. K., and Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315.

Liu, Y., Zhang, W., Zhang, K., You, Q., Yan, H., Jiao, Y., Jiang, J., Xu, W., and Su, Z. (2017). Genome-wide mapping of DNase I hypersensitive sites reveals chromatin accessibility changes in Arabidopsis euchromatin and heterochromatin regions under extended darkness. *Scientific Reports*, 7.

Lu, S. and Cullen, B. R. (2003). Analysis of the stimulatory effect of splicing on mRNA production and utilization in mammalian cells. *RNA*, 9(5):618–630.

Luo, M., Liu, X., Singh, P., Cui, Y., Zimmerli, L., and Wu, K. (2012). Chromatin modifications and remodeling in plant abiotic stress responses. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(2):129–136.

Mardis, E. R. (2010). The $1,000 genome, the $100,000 analysis? *Genome Medicine*, 2(11):84.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517.

Mavrich, T. N., Ioshikhes, I. P., Venters, B. J., Jiang, C., Tomsho, L. P., Qi, J., Schuster, S. C., Albert, I., and Pugh, B. F. (2008). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research*, 18(7):1073–1083.

Mayer, B. F., Ali-Benali, M. A., Demone, J., Bertrand, A., and Charron, J.-B. (2015). Cold acclimation induces distinctive changes in the chromatin state and transcript levels of COR genes in *Cannabis sativa* varieties with contrasting cold acclimation capacities. *Physiologia Plantarum*, 155(3):281–295.

Medina, J., Catalá, R., and Salinas, J. (2011). The CBFs: three Arabidopsis transcription factors to cold acclimate. *Plant Science*, 180(1):3–11.

Mercer, T. R., Edwards, S. L., Clark, M. B., Neph, S. J., Wang, H., Stergachis, A. B., John, S., Sandstrom, R., Li, G., Sandhu, K. S., et al. (2013). DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. *Nature Genetics*, 45(8):852–859.

Meshi, T. and Iwabuchi, M. (1995). Plant transcription factors. *Plant and Cell Physiology*, 36(8):1405–1420.

Mito, Y., Henikoff, J. G., and Henikoff, S. (2007). Histone replacement marks the boundaries of cis-regulatory domains. *Science*, 315(5817):1408–1411.

Murashige, T. and Skoog, F. (1962). A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiologia Plantarum*, 15(3):473–497.

Muse, G. W., Gilchrist, D. A., Nechaev, S., Shah, R., Parker, J. S., Grissom, S. F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. *Nature Genetics*, 39(12):1507–1511.

Natarajan, A., Yardımcı, G. G., Sheffield, N. C., Crawford, G. E., and Ohler, U. (2012). Predicting cell-type–specific gene expression from regions of open chromatin. *Genome Research*, 22(9):1711–1722.

Ogle, D. H. (2017). *FSA: Fisheries Stock Analysis*. R package version 0.8.13.

Ooi, S. K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.-P., Allis, C. D., et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to *de novo* methylation of DNA. *Nature*, 448(7154):714.

Oono, Y., Seki, M., Satou, M., Iida, K., Akiyama, K., Sakurai, T., Fujita, M., Yamaguchi-Shinozaki, K., and Shinozaki, K. (2006). Monitoring expression profiles of Arabidopsis genes during cold acclimation and deacclimation using DNA microarrays. *Functional & Integrative Genomics*, 6(3):212–234.

Pabo, C. O. and Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annual Review of Biochemistry*, 61(1):1053–1095.

Pajoro, A., Madrigal, P., Muiño, J. M., Matus, J. T., Jin, J., Mecchia, M. A., Debernardi, J. M., Palatnik, J. F., Balazadeh, S., Arif, M., et al. (2014). Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biology*, 15(3):R41.

Phelan, M. L., Schnitzler, G. R., and Kingston, R. E. (2000). Octamer transfer and creation of stably remodeled nucleosomes by human SWI-SNF and its isolated ATPases. *Molecular and Cellular Biology*, 20(17):6380–6389.

Prokhortchouk, E. and Defossez, P.-A. (2008). The cell biology of DNA methylation in mammals. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1783(11):2167–2173.

Pugh, B. F. and Tjian, R. (1991). Transcription from a TATA-less promoter requires a multisubunit TFIID complex. *Genes & Development*, 5(11):1935–1945.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Radman-Livaja, M. and Rando, O. J. (2010). Nucleosome positioning: how is it established, and why does it matter? *Developmental Biology*, 339(2):258–266.

Rando, O. J. (2007). Global patterns of histone modifications. *Current Opinion in Genetics & Development*, 17(2):94–99.

Raser, J. M. and O'Shea, E. K. (2004). Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–1814.

Reeves, R. and Nissen, M. S. (1990). The AT-DNA-binding domain of mammalian high mobility group I chromosomal proteins. A novel peptide motif for recognizing DNA structure. *Journal of Biological Chemistry*, 265(15):8573–8582.

Riechmann, J. L. (2002). Transcriptional regulation: a genomic overview. *The Arabidopsis Book*, page e0085.

Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C.-Z., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O., Samaha, R., et al. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, 290(5499):2105–2110.

Rose, A. B. (2002). Requirements for intron-mediated enhancement of gene expression in Arabidopsis. *RNA*, 8(11):1444–1453.

Rose, A. B., Carter, A., Korf, I., and Kojima, N. (2016). Intron sequences that stimulate gene expression in Arabidopsis. *Plant Molecular Biology*, 92(3):337–346.

Rymen, B., Fiorani, F., Kartal, F., Vandepoele, K., Inzé, D., and Beemster, G. T. (2007). Cold nights impair leaf growth and cell cycle progression in maize through transcriptional changes of cell cycle genes. *Plant Physiology*, 143(3):1429–1438.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., and Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome Biology*, 12(8):125.

Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898.

Seki, M., Narusaka, M., Ishida, J., Nanjo, T., Fujita, M., Oono, Y., Kamiya, A., Nakajima, M., Enju, A., Sakurai, T., et al. (2002). Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *The Plant Journal*, 31(3):279–292.

Shinwari, Z. K., Nakashima, K., Miura, S., Kasuga, M., Seki, M., Yamaguchi-Shinozaki, K., and Shinozaki, K. (1998). An Arabidopsis gene family encoding DRE/CRT binding proteins involved in low-temperature-responsive gene expression. *Biochemical and Biophysical Research Communications*, 250(1):161–170.

Shogren-Knaak, M., Ishii, H., Sun, J.-M., Pazin, M. J., Davie, J. R., and Peterson, C. L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science*, 311(5762):844–847.

Shu, H., Gruissem, W., and Hennig, L. (2013). Measuring Arabidopsis chromatin accessibility using DNase I-polymerase chain reaction and DNase I-chip assays. *Plant Physiology*, 162(4):1794–1801.

Siegfried, Z., Eden, S., Mendelsohn, M., Feng, X., Tsuberi, B.-Z., and Cedar, H. (1999). DNA methylation represses transcription in vivo. *Nature Genetics*, 22(2).

Singh, R. (2012). Climate change and food security. *Improving Crop Productivity in Sustainable Agriculture*, pages 1–22.

Sinha, S., Kukreja, B., Arora, P., Sharma, M., Pandey, G. K., Agarwal, M., and Chinnusamy, V. (2015). The omics of cold stress responses in plants. In *Elucidation of Abiotic Stress Signaling in Plants*, pages 143–194. Springer.

Sionit, N., Strain, B., and Flint, E. (1987). Interaction of temperature and CO2 enrichment on soybean: growth and dry matter partitioning. *Canadian Journal of Plant Science*, 67(1):59–67.

Smaczniak, C., Immink, R. G., Muiño, J. M., Blanvillain, R., Busscher, M., Busscher-Lange, J., Dinh, Q. P., Liu, S., Westphal, A. H., Boeren, S., et al. (2012). Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. *Proceedings of the National Academy of Sciences*, 109(5):1560–1565.

Song, L. and Crawford, G. E. (2010). Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2):pdb–prot5384.

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B.-K., Sheffield, N. C., Gräf, S., Huss, M., Keefe, D., et al. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research*, 21(10):1757–1767.

Spelt, C., Quattrocchio, F., Mol, J. N., and Koes, R. (2000). anthocyanin1 of petunia encodes a basic helix-loop-helix protein that directly activates transcription of structural anthocyanin genes. *The Plant Cell*, 12(9):1619–1631.

Steponkus, P. and Webb, M. (1992). Freeze-induced dehydration and membrane destabilization in plants. In *Water and Life*, pages 338–362. Springer.

Steponkus, P. L., Uemura, M., Joseph, R. A., Gilmour, S. J., and Thomashow, M. F. (1998). Mode of action of the COR15a gene on the freezing tolerance of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*, 95(24):14570–14575.

Stergachis, A. B., Neph, S., Reynolds, A., Humbert, R., Miller, B., Paige, S. L., Vernot, B., Cheng, J. B., Thurman, R. E., Sandstrom, R., et al. (2013). Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell*, 154(4):888–903.

Stockinger, E. J., Gilmour, S. J., and Thomashow, M. F. (1997). *Arabidopsis thaliana* CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. *Proceedings of the National Academy of Sciences*, 94(3):1035–1040.

Stünkel, W., Kober, I., and Seifart, K. H. (1997). A nucleosome positioned in the distal promoter region activates transcription of the human U6 gene. *Molecular and Cellular Biology*, 17(8):4397–4405.

Sullivan, A. M., Arsovski, A. A., Lempe, J., Bubb, K. L., Weirauch, M. T., Sabo, P. J., Sandstrom, R., Thurman, R. E., Neph, S., Reynolds, A. P., et al. (2014). Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Reports*, 8(6):2015–2030.

Sullivan, A. M., Bubb, K. L., Sandstrom, R., Stamatoyannopoulos, J. A., and Queitsch, C. (2015). DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants. *Current Plant Biology*, 3:40–47.

Suzuki, M. M. and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465–476.

Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D., and Patel, D. J. (2007). How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nature Structural & Molecular Biology*, 14(11):1025.

Team, R. C. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Thakur, P., Kumar, S., Malik, J. A., Berger, J. D., and Nayyar, H. (2010). Cold stress effects on reproductive development in grain crops: an overview. *Environmental and Experimental Botany*, 67(3):429–443.

Thanos, D. and Maniatis, T. (1992). The high mobility group protein HMG I (Y) is required for NF-κB-dependent virus induction of the human IFN-β gene. *Cell*, 71(5):777–789.

Thomashow, M. F. (2010). Molecular basis of plant cold acclimation: insights gained from studying the CBF cold response pathway. *Plant Physiology*, 154(2):571–577.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Research*, 25(10):1491–1498.

Van Holde, K. E. (2012). *Chromatin*. Springer Science & Business Media.

Vavouri, T. and Elgar, G. (2005). Prediction of cis-regulatory elements using binding site matricesthe successes, the failures and the reasons for both. *Current Opinion in Genetics & Development*, 15(4):395–402.

Vierstra, J., Wang, H., John, S., Sandstrom, R., and Stamatoyannopoulos, J. A. (2014). Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nature Methods*, 11(1):66–72.

Vining, K. J., Pomraning, K. R., Wilhelm, L. J., Priest, H. D., Pellegrini, M., Mockler, T. C., Freitag, M., and Strauss, S. H. (2012). Dynamic DNA cytosine methylation in the Populus trichocarpa genome: tissue-level variation and relationship to gene expression. *BMC Genomics*, 13(1):27.

Wade, P. A., Gegonne, A., Jones, P. L., Ballestar, E., Aubry, F., and Wolffe, A. P. (1999). Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nature Genetics*, 23(1).

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

Wanner, L. A. and Junttila, O. (1999). Cold-induced freezing tolerance in Arabidopsis. *Plant Physiology*, 120(2):391–400.

Weber, A. P., Weber, K. L., Carr, K., Wilkerson, C., and Ohlrogge, J. B. (2007). Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiology*, 144(1):32–42.

Weirauch, M. T. and Hughes, T. (2011). A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. In *A handbook of transcription factors*, pages 25–73. Springer.

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443.

Wiench, M., John, S., Baek, S., Johnson, T. A., Sung, M.-H., Escobar, T., Simmons, C. A., Pearce, K. H., Biddie, S. C., Sabo, P. J., et al. (2011). DNA methylation status predicts cell type-specific enhancer activity. *The EMBO Journal*, 30(15):3028–3039.

Workman, J. and Kingston, R. (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annual Review of Biochemistry*, 67(1):545–579.

Workman, J. L. and Kingston, R. E. (1992). Nucleosome core displacement in vitro via a metastable transcription factor-nucleosome complex. *Science*, 258(5089):1780–1785.

Wu, C. (1980). The 5'ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature*, 286(5776):854–860.

Wu, C., Wong, Y.-C., and Elgin, S. C. (1979). The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell*, 16(4):807–814.

Wu, Y., Zhang, W., and Jiang, J. (2014). Genome-wide nucleosome positioning is orchestrated by genomic regions associated with DNase I hypersensitivity in rice. *PLoS Genetics*, 10(5):e1004378.

Xiong, L., Ishitani, M., and Zhu, J.-K. (1999). Interaction of osmotic stress, temperature, and abscisic acid in the regulation of gene expression in Arabidopsis. *Plant Physiology*, 119(1):205–212.

Yamaguchi-Shinozaki, K. and Shinozaki, K. (1994). A novel cis-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. *The Plant Cell*, 6(2):251–264.

Yuan, G.-C., Liu, Y.-J., Dion, M. F., Slack, M. D., Wu, L. F., Altschuler, S. J., and Rando, O. J. (2005). Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*, 309(5734):626–630.

Yudkovsky, N., Logie, C., Hahn, S., and Peterson, C. L. (1999). Recruitment of the SWI/SNF chromatin remodeling complex by transcriptional activators. *Genes and Development*, 13(18):2369–2374.

Yun, J., Kim, Y.-S., Jung, J.-H., Seo, P. J., and Park, C.-M. (2012). The AT-hook motif-containing protein AHL22 regulates flowering initiation by modifying FLOWERING LOCUS T chromatin in Arabidopsis. *Journal of Biological Chemistry*, 287(19):15307–15316.

Zaret, K. S. and Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes & Development*, 25(21):2227–2241.

Zarka, D. G., Vogel, J. T., Cook, D., and Thomashow, M. F. (2003). Cold induction of Arabidopsis *CBF* genes involves multiple ICE (inducer of *CBF* expression) promoter elements and a cold-regulatory circuit that is desensitized by low temperature. *Plant Physiology*, 133(2):910–918.

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142.

Zhang, W., Wu, Y., Schnable, J. C., Zeng, Z., Freeling, M., Crawford, G. E., and Jiang, J. (2012a). High-resolution mapping of open chromatin in the rice genome. *Genome Research*, 22(1):151–162.

Zhang, W., Zhang, T., Wu, Y., and Jiang, J. (2012b). Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *The Plant Cell*, 24(7):2719–2731.

Zhang, W., Zhang, T., Wu, Y., and Jiang, J. (2014). Open chromatin in plant genomes. *Cytogenetic and Genome Research*, 143(1-3):18–27.

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y. V., Pellegrini, M., Goodrich, J., and Jacobsen, S. E. (2007). Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS Biology*, 5(5):e129.

Zhang, X., Henriques, R., Lin, S.-S., Niu, Q.-W., and Chua, N.-H. (2006). Agrobacterium-mediated transformation of *Arabidopsis thaliana* using the floral dip method. *Nature Protocols*, 1(2):641–646.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS One*, 9(1):e78644.

Zhou, M., Shen, C., Wu, L., Tang, K., and Lin, J. (2011). CBF-dependent signaling pathway: a key responder to low temperature stress in plants. *Critical Reviews in Biotechnology*, 31(2):186–192.

Zhu, B., Zhang, W., Zhang, T., Liu, B., and Jiang, J. (2015). Genome-wide prediction and validation of intergenic enhancers in Arabidopsis using open chromatin signatures. *The Plant Cell*, 27(9):2415–2426.

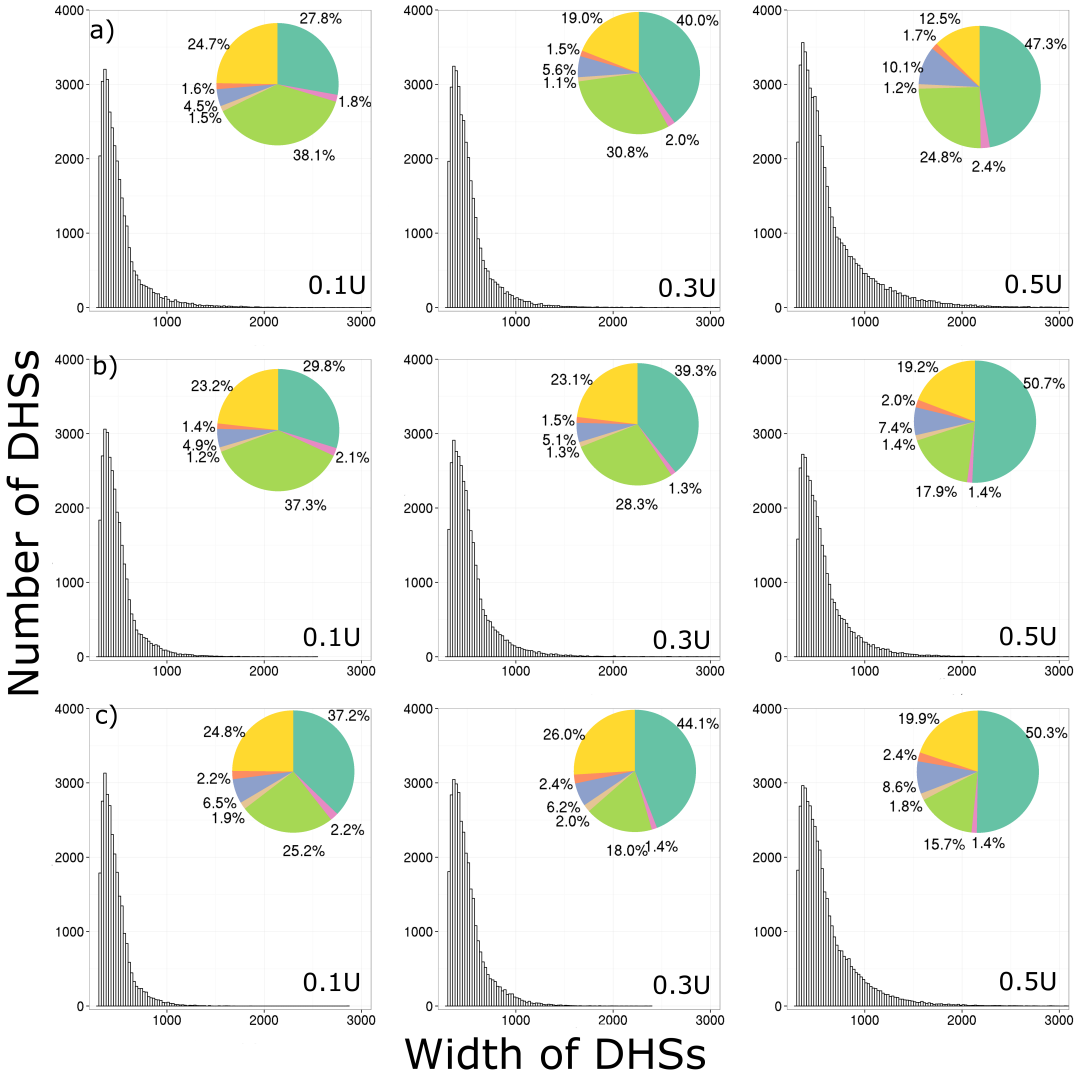
Zhu, J., Jeong, J. C., Zhu, Y., Sokolchik, I., Miyazaki, S., Zhu, J.-K., Hasegawa, P. M., Bohnert, H. J., Shi, H., Yun, D.-J., et al. (2008). Involvement of Arabidopsis HOS15 in histone deacetylation and cold tolerance. *Proceedings of the National Academy of Sciences*, 105(12):4945–4950.

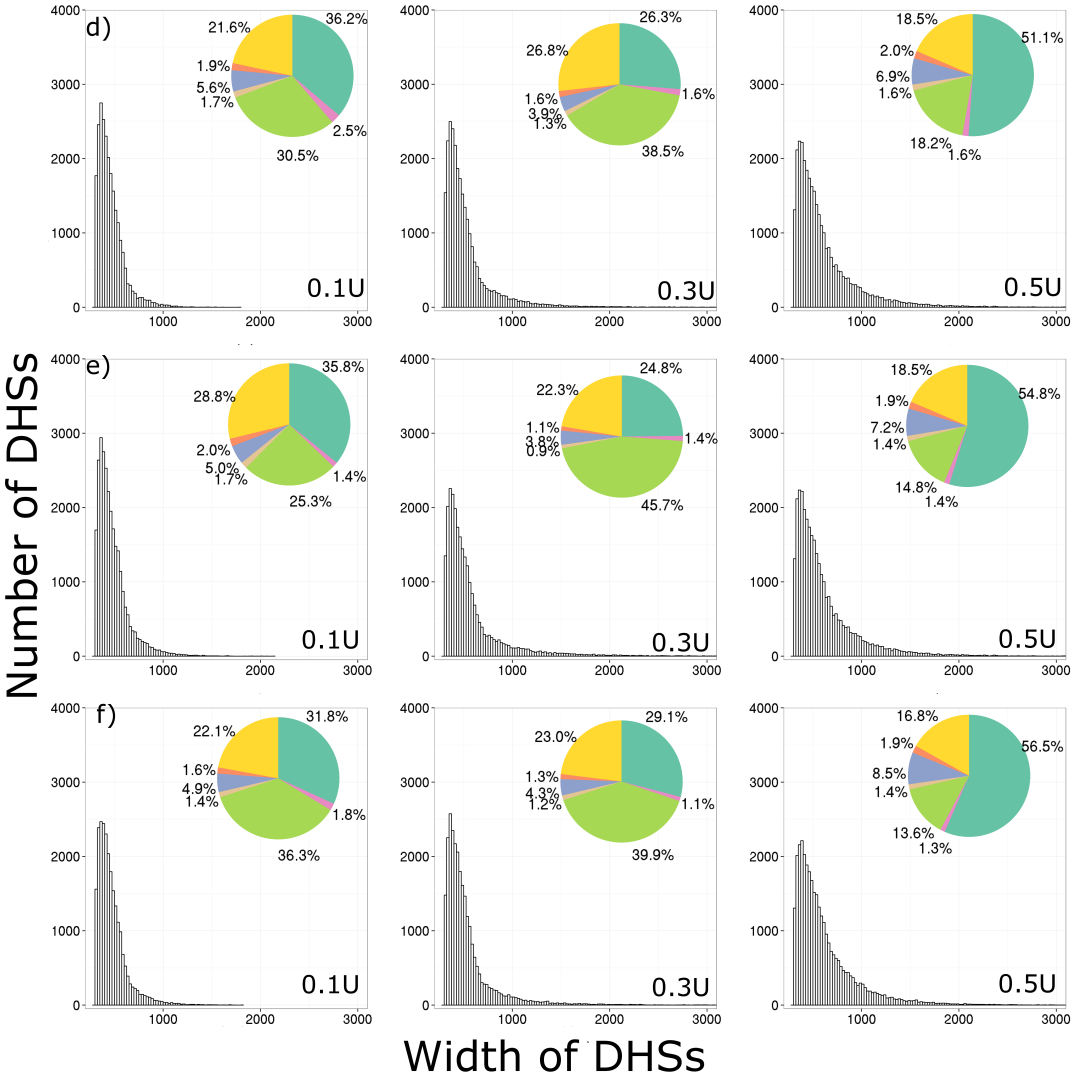
Zhu, W., Hu, B., Becker, C., Doğan, E. S., Berendzen, K. W., Weigel, D., and Liu, C. (2017). Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific Arabidopsis hybrid. *Genome Biology*, 18(1):157.

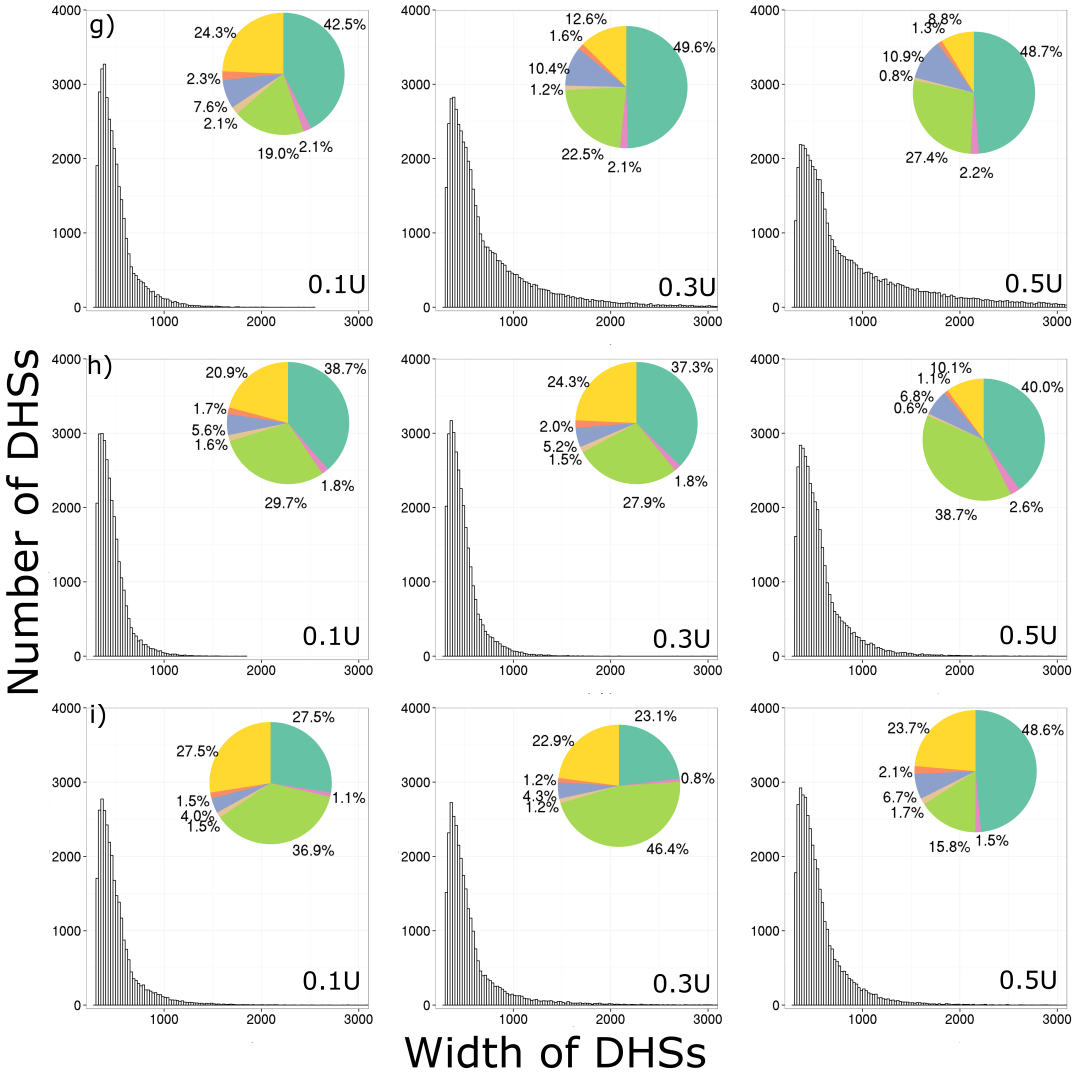
Zou, C., Sun, K., Mackaluso, J. D., Seddon, A. E., Jin, R., Thomashow, M. F., and Shiu, S.-H. (2011). Cis-regulatory code of stress-responsive transcription in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*, 108(36):14992–14997.

Appendix A

Replicate DNase Digestion Profiles




d)
21.6%
36.2%
1.9%
5.6%
1.7%
2.5%
30.5%
0.1U
26.3%
26.8%
1.6%
1.6%
3.9%
1.3%
38.5%
0.3U
18.5%
51.1%
2.0%
6.9%
1.6%
18.2%
1.6%
0.5U
e)
35.8%
28.8%
2.0%
5.0%
1.7%
1.4%
25.3%
0.1U
24.8%
22.3%
1.1%
3.8%
0.9%
1.4%
45.7%
0.3U
18.5%
54.8%
1.9%
7.2%
1.4%
14.8%
1.4%
0.5U
f)
31.8%
22.1%
1.6%
4.9%
1.4%
1.8%
36.3%
0.1U
29.1%
23.0%
1.3%
4.3%
1.2%
1.1%
39.9%
0.3U
16.8%
56.5%
1.9%
8.5%
1.4%
13.6%
1.3%
0.5U
Number of DHSs
Width of DHSs


g)
h)
i)
0.1U
0.3U
0.5U
Number of DHSs
Width of DHSs
0
1000
2000
3000
4000
24.3%
42.5%
2.3%
7.6%
2.1%
19.0%
2.1%
12.6%
1.6%
49.6%
10.4%
1.2%
22.5%
2.1%
8.8%
1.3%
48.7%
10.9%
0.8%
27.4%
2.2%
20.9%
38.7%
1.7%
5.6%
1.6%
29.7%
1.8%
24.3%
37.3%
2.0%
5.2%
1.5%
27.9%
1.8%
10.1%
1.1%
40.0%
6.8%
0.6%
38.7%
2.6%
27.5%
27.5%
1.1%
1.5%
4.0%
1.5%
36.9%
23.1%
22.9%
0.8%
1.2%
4.3%
1.2%
46.4%
23.7%
48.6%
2.1%
6.7%
1.7%
15.8%
1.5%

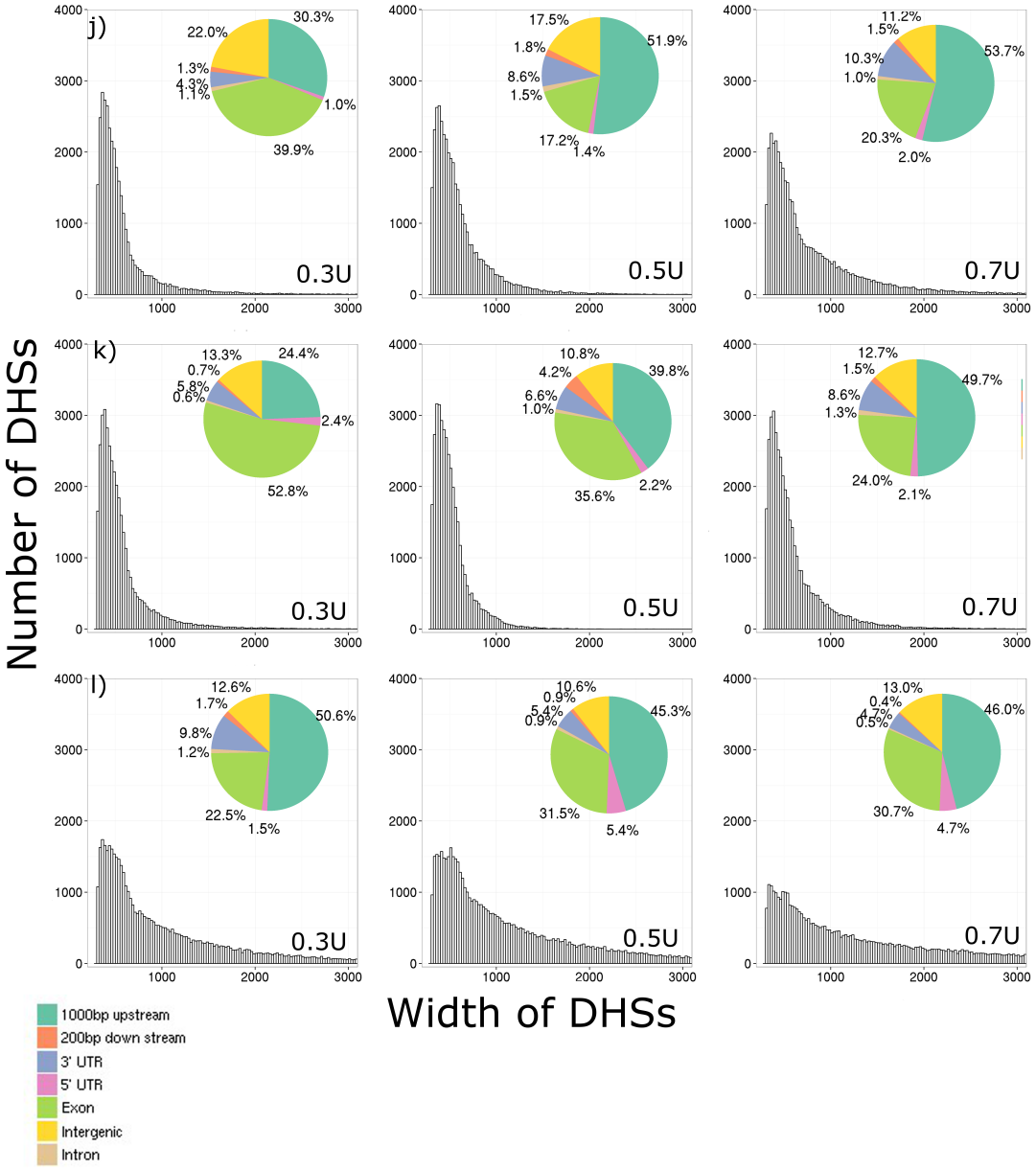


Figure A.1: **DNase digestion profiles of individual replicates.** DNase-DTS bar plots displaying the frequency of DHSs with respect to DHS size. Also displayed are pie graphs showing the distribution of DHS within genomic locations. DHSs were classified into genomic categories in order of importance, upstream 1000 bp, 5'UTR, 3'UTR, exon, intron, downstream 200 bp, and intergenic. Concentration of DNase shown in lower right of each subplot. a) Epidermal replicate #1 b) Epidermal replicate #2. c) Epidermal replicate #3. d) Endodermal replicate #1. e) Endodermal replicate #2. f) Endodermal replicate #3 . g) Epidermal cold replicate #1. h) Epidermal cold replicate #2. i) Epidermal cold replicate #3. j) Endodermal cold replicate #1. k) Endodermal cold replicate #2. l) Endodermal cold replicate #3.

Appendix B

DE/DA genes

B.1 List of epidermal DE/DA genes

Table B.1: List of epidermal DE/DA genes

AT1G05562	Unknown
AT1G12070	Immunoglobulin E-set superfamily protein (PTHR10980:SF30)
AT1G13480	SUBFAMILY NOT NAMED (PTHR31050:SF2)
AT1G21130	Caffeic acid 3-O-methyltransferase-like protein-related (PTHR11746:SF118)
AT1G23149	Unknown
AT1G23150	Gb—AAC00605.1 (PTHR33270:SF14)
AT1G30530	UDP-glycosyltransferase 78D1-related (PTHR11926:SF451)
AT1G37130	Nitrate reductase [NADH] 1-related (PTHR19370:SF182)
AT1G44100	Amino acid permease 5 (PTHR22950:SF350)
AT1G54030	GDSL esterase/lipase 1-related (PTHR22835:SF485)
AT1G64300	Protein kinase (PTHR44630:SF3)
AT1G64940	Cytochrome P450 89A2-related (PTHR24298:SF45)
AT1G66200	Glutamine synthetase cytosolic isozyme 1-2 (PTHR20852:SF58)

AT2G04090 MATE efflux family protein-related (PTHR11206:SF189)

AT2G15830 SUBFAMILY NOT NAMED (PTHR37705:SF1)

AT2G19800 Inositol oxygenase 2-related (PTHR12588:SF1)

AT2G21188 Unknown

AT2G28760 UDP-glucuronic acid decarboxylase 3-related (PTHR43078:SF9)

AT2G43610 Chitinase family protein-related (PTHR22595:SF104)

AT2G45220 pectinesterase/pectinesterase inhibitor 17-related (PTHR31707:SF11)

AT2G45400 Protein BRI1-5 ENHANCED 1 (PTHR10366:SF560)

AT3G03640 Beta-D-glucopyranosyl abscisate beta-glucosidase-related (PTHR10353:SF81)

AT3G03650 Exostosin family protein (PTHR11062:SF51)

AT3G06210 ARM repeat superfamily protein-related (PTHR33115:SF3)

AT3G07410 Ras-related protein RABA5b (PTHR24073:SF884)

AT3G13790 Beta-fructofuranosidase, insoluble isoenzyme CWINV1-related (PTHR31953:SF16)

AT3G16420 Jacalin-like lectin domain-containing protein-related (PTHR23244:SF208)

AT3G20935 Cytochrome P450-related (PTHR24298:SF59)

AT3G20960 Cytochrome P450-related (PTHR24298:SF59)

AT3G52820 Purple acid phosphatase 21-related (PTHR22953:SF7)

AT3G56060 Glucose-methanol-choline (GMC) oxidoreductase family protein (PTHR11552:SF141)

AT3G59480 fructokinase-1-related (PTHR43085:SF6)

AT4G00730 *C. elegans* Homeobox (PTHR24326:SF511)

AT4G03480 Ankyrin repeat family protein-related (PTHR24177:SF109)

AT4G12910 Serine carboxypeptidase-like 20-related (PTHR11802:SF188)

AT4G13990 xyloglucan galactosyltransferase GT14-related (PTHR11062:SF110)

AT4G15215 ABC transporter G family member 30-related (PTHR19241:SF449)

AT4G15760 FAD/NAD(P)-binding oxidoreductase family protein-related (PTHR13789:SF1)

AT4G24670 Tryptophan aminotransferase-related protein 2 (PTHR43795:SF22)

AT4G25310	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein-related (PTHR10
AT4G28940	Phosphorylase superfamily protein (PTHR21234:SF19)
AT4G31320	Auxin induced like-protein (PTHR31374:SF59)
AT4G35840	NEP1-interacting protein 1-related (PTHR14155:SF307)
AT4G37160	Pectinesterase like protein-related (PTHR11709:SF113)
AT4G37330	Cytochrome P450 81D1-related (PTHR24298:SF205)
AT5G24070	peroxidase 26-related (PTHR31235:SF78)
AT5G25170	Expressed protein (PTHR12378:SF12)
AT5G33290	Xylogalacturonan beta-1,3-xylosyltransferase (PTHR11062:SF61)
AT5G41315	Transcription factor EGL1-related (PTHR11514:SF28)
AT5G41800	GABA transporter 2-related (PTHR22950:SF229)
AT5G43030	C1 domain-containing protein-related (PTHR32410:SF152)
AT5G43040	C1 domain-containing protein-related (PTHR32410:SF152)
AT5G45280	Pectin acetylesterase 11-related (PTHR21562:SF37)
AT5G56870	Beta-galactosidase 12-related (PTHR23421:SF63)

B.2 List of endodermal DE/DA genes

Table B.2: List of endodermal DE/DA genes

AT1G01120	3-ketoacyl-CoA synthase 1-related (PTHR31561:SF61)
AT1G03440	Leucine-rich repeat family protein-related (PTHR44450:SF2)
AT1G14040	Phosphate transporter PHO1 homolog 2-related (PTHR10783:SF43)
AT1G17970	RING zinc finger protein— 69105-67310-related (PTHR22937:SF98)
AT1G19450	Sugar transporter ERD6-like 4-related (PTHR23500:SF337)
AT1G21520	Unknown
AT1G22220	SUBFAMILY NOT NAMED (PTHR31215:SF6)

AT1G22470 SUBFAMILY NOT NAMED (PTHR34046:SF3)

AT1G23050 Hydroxyproline-rich glycoprotein family protein (PTHR37702:SF1)

AT1G43910 AAA-type ATPase family protein-related (PTHR23070:SF31)

AT1G46264 Heat stress transcription factor B-4 (PTHR10015:SF159)

AT1G48480 inactive receptor kinase RLK902-related (PTHR27008:SF19)

AT1G53440 SUBFAMILY NOT NAMED (PTHR27006:SF93)

AT1G61590 SUBFAMILY NOT NAMED (PTHR27001:SF495)

AT1G64330 Protein NETWORKED 3A-related (PTHR32258:SF2)

AT1G65710 SUBFAMILY NOT NAMED (PTHR34367:SF1)

AT1G67050 F1O19.11 protein (PTHR31722:SF10)

AT1G69295 Unknown

AT1G70580 Glutamate–glyoxylate aminotransferase 1-related (PTHR11751:SF373)

AT1G70640 PB1_UP2 domain-containing protein (PTHR31066:SF1)

AT1G72230 SUBFAMILY NOT NAMED (PTHR33021:SF151)

AT1G76700 Chaperone protein dnaJ 10-related (PTHR44094:SF3)

AT2G16970 Hippocampus abundant transcript 1 protein (PTHR23504:SF1)

AT2G22000 Unknown

AT2G25150 HXXXD-type acyl-transferase family protein-related (PTHR31147:SF25)

AT2G28670 Dirigent protein 10-related (PTHR21495:SF41)

AT2G28790 SUBFAMILY NOT NAMED (PTHR31048:SF79)

AT2G31360 Delta-9 acyl-lipid desaturase 1-related (PTHR11351:SF2)

AT2G36290 Alpha/beta-Hydrolases superfamily protein-related (PTHR10992:SF997)

AT2G41050 Lysosomal amino acid transporter 1 (PTHR16201:SF34)

AT2G46590 Dof zinc finger protein DOF1.2-related (PTHR31992:SF15)

AT2G46680 Homeobox-leucine zipper protein ATHB-12-related (PTHR24326:SF122)

AT3G13000 SUBFAMILY NOT NAMED (PTHR23054:SF10)

AT3G23160 SUBFAMILY NOT NAMED (PTHR31371:SF2)

AT3G24020 Dirigent protein 16-related (PTHR21495:SF86)

AT3G24030 Hydroxyethylthiazole kinase (PTHR44423:SF1)

AT3G28130 SUBFAMILY NOT NAMED (PTHR31218:SF64)

AT3G47220 1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase classes I and II (PTHR10336:SF3

AT3G50230 Leucine-rich repeat protein kinase family protein (PTHR27008:SF1)

AT3G59140 ABC transporter C family member 10 (PTHR24223:SF192)

AT4G15330 Cytochrome P450-related (PTHR24298:SF59)

AT4G18610 Protein LIGHT-DEPENDENT SHORT HYPOCOTYLS 9 (PTHR31165:SF8)

AT4G26140 Beta-galactosidase 12-related (PTHR23421:SF63)

AT4G26200 1-aminocyclopropane-1-carboxylate synthase 7 (PTHR43795:SF28)

AT4G26320 Arabinogalactan peptide 12-related (PTHR34114:SF2)

AT4G35550 Homeobox protein rough (PTHR24326:SF220)

AT4G36900 Ethylene-responsive transcription factor ERF008-related (PTHR31729:SF0)

AT4G40085 Unknown

AT5G01100 O-fucosyltransferase family protein (PTHR31741:SF4)

AT5G05220 SUBFAMILY NOT NAMED (PTHR37758:SF1)

AT5G07830 Heparanase-like protein 1-related (PTHR14363:SF21)

AT5G18690 Classical arabinogalactan protein 25 (PTHR35725:SF2)

AT5G19740 Glutamate carboxypeptidase 2 homolog (PTHR10404:SF46)

AT5G24380 Metal-nicotianamine transporter YSL2-related (PTHR31645:SF4)

AT5G41040 Trichothecene 3-O-acetyltransferase (PTHR31642:SF119)

AT5G41790 Centrosomin (PTHR13140:SF694)

AT5G42500 Dirigent protein 1-related (PTHR21495:SF71)

AT5G52790 SUBFAMILY NOT NAMED (PTHR12064:SF36)

AT5G57480 Mitochondrial chaperone BCS1 (PTHR23070:SF17)

AT5G62880	Rac-like GTP-binding protein ARAC10-related (PTHR24072:SF255)
AT5G63590	Flavonol synthase 3-related (PTHR10209:SF382)
AT5G65230	MYB transcription factor (PTHR10641:SF882)
AT5G65530	Protein kinase family protein-related (PTHR27001:SF105)
AT5G66390	Peroxidase 36-related (PTHR31388:SF3)

B.3 List of epidermal cold DE/DA genes

Table B.3: List of epidermal cold DE/DA genes

AT1G03940	Coumaroyl-CoA:anthocyanidin 3-O-glucoside-6”-O-coumaroyltransferase 1-related (PTHR3162
AT1G10370	Glutathione S-transferase U11-related (PTHR11260:SF487)
AT1G10410	CW14 protein (PTHR31558:SF3)
AT1G11330	G-type lectin S-receptor-like serine/threonine-protein kinase SD1-13 (PTHR27002:SF165)
AT1G24530	F21J9.19 (PTHR22844:SF199)
AT1G31020	Thioredoxin O1, mitochondrial-related (PTHR10438:SF275)
AT1G33600	Leucine-rich repeat (LRR) family protein-related (PTHR45271:SF1)
AT1G46768	Ethylene-responsive transcription factor ERF008-related (PTHR31729:SF0)
AT1G55820	GBF-interacting protein 1-like (PTHR12758:SF30)
AT1G63900	E3 ubiquitin-protein ligase SP1-related (PTHR12183:SF18)
AT1G70710	Endoglucanase 4-related (PTHR22298:SF28)
AT1G73480	Alpha/beta-Hydrolases superfamily protein (PTHR11614:SF114)
AT1G80130	F18B13.21 protein-related (PTHR26312:SF75)
AT1G80280	Alpha/beta-Hydrolases superfamily protein-related (PTHR43689:SF2)
AT2G04650	ADP-glucose pyrophosphorylase-like protein (PTHR22572:SF123)
AT2G15790	Peptidyl-prolyl cis-trans isomerase CYP26-1-related (PTHR11071:SF381)
AT2G21160	Translocon-associated protein subunit alpha (PTHR12924:SF0)

AT2G21580 40S ribosomal protein S25 (PTHR12850:SF5)

AT2G21620 Dessication responsive protein (PTHR31964:SF113)

AT2G22170 Expressed protein (PTHR31718:SF19)

AT2G23120 Expressed protein-related (PTHR33922:SF9)

AT2G28900 Outer envelope pore protein 16-1, chloroplastic (PTHR15371:SF2)

AT2G41190 Expressed protein (PTHR22950:SF358)

AT2G46680 Homeobox-leucine zipper protein ATHB-12-related (PTHR24326:SF122)

AT3G08630 Protein RETICULATA-RELATED 2, chloroplastic-related (PTHR31620:SF5)

AT3G14790 Trifunctional UDP-glucose 4,6-dehydratase (PTHR43000:SF8)

AT3G46460 Ubiquitin-conjugating enzyme E2 13-related (PTHR24067:SF237)

AT3G48890 Membrane steroid-binding protein 2 (PTHR10281:SF45)

AT3G52570 SUBFAMILY NOT NAMED (PTHR43248:SF6)

AT4G02520 Glutathione S-transferase F14-related (PTHR43900:SF7)

AT4G09020 Isoamylase 3, chloroplastic (PTHR43002:SF3)

AT4G25870 Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein (PTHR31042:SI

AT4G33070 Pyruvate decarboxylase 1-related (PTHR43452:SF1)

AT4G33140 SUBFAMILY NOT NAMED (PTHR35134:SF2)

AT4G38130 Histone deacetylase 19 (PTHR10625:SF178)

AT5G02590 Tetratricopeptide repeat domain-containing protein (PTHR26312:SF76)

AT5G06760 Late embryogenesis abundant protein 4-5 (PTHR33493:SF2)

AT5G14040 Mitochondrial phosphate carrier protein 3, mitochondrial (PTHR24089:SF460)

AT5G17220 Glutathione S-transferase F11-related (PTHR43900:SF18)

AT5G19440 Alcohol dehydrogenase-like protein-related (PTHR10366:SF564)

AT5G19550 Aspartate aminotransferase (PTHR11879:SF22)

AT5G20180 39S ribosomal protein L36, mitochondrial (PTHR18804:SF13)

AT5G20320 Dicer-like protein 4 (PTHR14950:SF15)

AT5G27760	Respiratory supercomplex factor 2, mitochondrial (PTHR28018:SF2)
AT5G28540	78 kDa glucose-regulated protein (PTHR19375:SF144)
AT5G50450	SUBFAMILY NOT NAMED (PTHR12298:SF37)
AT5G55180	O-Glycosyl hydrolases family 17 protein (PTHR32227:SF102)
AT5G55770	C1 domain-containing protein-related (PTHR32410:SF152)
AT5G57290	Unknown
AT5G61760	Inositol polyphosphate multikinase (PTHR12400:SF21)
AT5G66100	La-related protein 1B-related (PTHR22792:SF99)
AT5G66330	Leucine-rich repeat-containing protein (PTHR44632:SF2)
AT5G67470	Formin homology 2 domain containing ortholog, isoform I (PTHR23213:SF269)

B.4 List of endodermal cold DE/DA genes

Table B.4: List of endodermal cold DE/DA genes

AT1G09780	2,3-bisphosphoglycerate-independent phosphoglycerate mutase 1-related (PTHR31637:SF6)
AT1G10070	Branched-chain-amino-acid aminotransferase 1, mitochondrial-related (PTHR42825:SF5)
AT1G12200	Flavin-containing monooxygenase FMO GS-OX-like 1-related (PTHR23023:SF198)
AT1G14820	SUBFAMILY NOT NAMED (PTHR10174:SF121)
AT1G16470	Unknown
AT1G16900	Alpha-1,2-mannosyltransferase ALG9-related (PTHR22760:SF2)
AT1G17170	Glutathione S-transferase U19-related (PTHR11260:SF493)
AT1G18540	60S ribosomal protein L6 (PTHR10715:SF0)
AT1G24510	T-complex protein 1 subunit epsilon (PTHR11353:SF94)
AT1G46768	Ethylene-responsive transcription factor ERF008-related (PTHR31729:SF0)
AT1G47710	Accessory gland protein Acp76A-related (PTHR11461:SF211)
AT1G48850	Chorismate synthase (PTHR21085:SF0)

AT1G55820 GBF-interacting protein 1-like (PTHR12758:SF30)

AT1G55915 SUBFAMILY NOT NAMED (PTHR23153:SF37)

AT1G62570 Flavin-containing monooxygenase FMO GS-OX-like 1-related (PTHR23023:SF198)

AT1G71750 Hypoxanthine PhosphoRibosylTransferase homolog (PTHR43340:SF1)

AT1G76520 Protein PIN-LIKES 1-related (PTHR31651:SF5)

AT2G04650 ADP-glucose pyrophosphorylase-like protein (PTHR22572:SF123)

AT2G17120 LysM domain-containing GPI-anchored protein 2 (PTHR33734:SF5)

AT2G28550 AP2-like ethylene-responsive transcription factor SMZ-related (PTHR32467:SF46)

AT2G35840 sucrose-phosphatase 1-related (PTHR12526:SF2)

AT2G36010 Transcription factor E2F/dimerisation partner (TDP) family protein (PTHR12081:SF18)

AT2G36620 60S ribosomal protein L24-1-related (PTHR10792:SF18)

AT2G36930 C2H2-type zinc finger-containing protein (PTHR20863:SF54)

AT2G38730 Peptidyl-prolyl cis-trans isomerase H (PTHR11071:SF58)

AT2G46390 Unknown

AT3G02420 SUBFAMILY NOT NAMED (PTHR30603:SF18)

AT3G03640 Beta-D-glucopyranosyl abscisate beta-glucosidase-related (PTHR10353:SF81)

AT3G08640 Protein RETICULATA-RELATED 2, chloroplastic-related (PTHR31620:SF5)

AT3G13570 Serine/arginine-rich SC35-like splicing factor SCL30A (PTHR23147:SF41)

AT3G14890 Bifunctional polynucleotide phosphatase/kinase (PTHR12083:SF9)

AT3G25585 FI05338p (PTHR10414:SF37)

AT3G53500 Serine/arginine-rich splicing factor RS2Z32-related (PTHR23147:SF21)

AT3G53880 Aldo-keto reductase family 4 member C11-related (PTHR11732:SF365)

AT3G55610 Delta-1-pyrroline-5-carboxylate synthase (PTHR11063:SF8)

AT3G59380 Protein farnesyltransferase/geranylgeranyltransferase type-1 subunit alpha (PTHR11129:SF1

AT4G02520 Glutathione S-transferase F14-related (PTHR43900:SF7)

AT4G13180 SUBFAMILY NOT NAMED (PTHR43361:SF1)

AT4G14320	IP17351p (PTHR10369:SF3)
AT4G14800	Proteasome subunit beta type-2-A-related (PTHR11599:SF99)
AT4G19460	Glycosyltransferase (PTHR12526:SF316)
AT4G24820	26S proteasome non-ATPase regulatory subunit 6 (PTHR14145:SF1)
AT4G26120	Unknown
AT4G31810	3-hydroxyisobutyryl-CoA hydrolase-like protein 2, mitochondrial (PTHR43176:SF8)
AT5G11430	SPOC and transcription elongation factor S-II domain protein-related (PTHR11477:SF20)
AT5G11640	Protein TMX2-CTNND1-related (PTHR15853:SF0)
AT5G20360	Octicosapeptide/Phox/Bem1p and tetratricopeptide repeat domain-containing protein (PTHR229
AT5G27760	Respiratory supercomplex factor 2, mitochondrial (PTHR28018:SF2)
AT5G44390	Berberine bridge enzyme-related (PTHR32448:SF38)
AT5G47550	Cysteine proteinase inhibitor 5 (PTHR11413:SF65)
AT5G52390	PAR1 protein (PTHR33649:SF2)
AT5G57950	26S proteasome non-ATPase regulatory subunit 9 (PTHR12651:SF1)
AT5G60590	YrdC domain-containing protein, mitochondrial (PTHR17490:SF10)

Curriculum Vitae
Shawn Hoogstra

EDUCATION

Bachelor of Science, Honors Specialization: Genetics
Western University, London, Ontario, 2011 -- 2015

Master of Science, Biology: Bioinformatics
Western University, London, Ontario, Expected Fall 2017

Thesis title: "Chromatin accessibility dynamics in the Arabidopsis root epidermis and endodermis during cold acclimation"

TEACHING/SUPERVISING EXPERIENCE

Teaching Assistant, Organismal Physiology
Biology Department, University of Western (2015--2017)

- Prepared an organismal physiology lab, explained background material for each lab, supervised students performing various experiments, and graded lab reports

Teaching Assistant, Genetics
Biology Department, University of Western (2015--2017)

- Taught a genetics tutorial explaining additional lecture material, genetic techniques, and genetic processes using planned lessons and assignments

HONORS AND AWARDS

Dean's Honor List, University of Ottawa 2011--2012
Dean's Honor List, University of Western 2012--2015
Hellen I. Battle Scholarship, University of Western 2013--2014

- Awarded to the highest and second highest ranking student in the third year of the BSc Biology program

RESEARCH EXPERIENCE

Graduate Research Assistant September 2015 -- Present
Department of Biology, University of Western Ontario
& London Research Development Centre
September 2015 -- Present: Master's Student, Western University, PI: Dr. Ryan Austin

- Understand cell-type specific chromatin accessibility changes in Arabidopsis under cold stress using bioinformatics and next generation sequencing technologies such Nextera (Illumina) and the Illumina Miseq.
- Utilized Python, R, Java, Unix, Sed, Awk, & Bash for the development of novel software for next generation sequencing and metabolomics analyses
- Computational analyze and integrate various 'big' data including RNA-seq, DNase-seq, ChIP-seq, Methyl-seq, microarrays, motif mapping, & metabolomics data

POSTER PRESENTATIONS

Hoogstra, S., Leckie, K., Austin, R. Transcriptional regulation of cell-type specific expression in the *Arabidopsis* root. (2017, July). Synthetic Biology Symposium. University of Western, Ontario.

TECHNICAL SKILLS

Programming languages: Python, Java, R, C, Unix, Sed, Awk, Bash
Operating systems: Linux (Ubuntu), Windows OS, Mac OS